PRCA-99031-US
09/472,249

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: １９９９年　１月２９日

出願番号
Application Number: 平成１１年特許願第０２２９１５号

出願人
Applicant (s): 株式会社リコー

１９９９年１０月２２日

特許庁長官
Commissioner,
Patent Office 近藤隆彦

出証番号　出証特平１１－３０７２８０８

【書類名】　特許願

【整理番号】　9806050

【提出日】　平成11年 1月29日

【あて先】　特許庁長官殿

【国際特許分類】　G06F　3/00

【発明の名称】　文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【請求項の数】　15

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　嶋田　敦夫

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　宮地　達生

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　剣持　栄治

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　山崎　真湖人

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　武谷　一寿

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　長束　哲郎

【特許出願人】

【識別番号】　000006747

出証特平１１－３０７２８０８

【氏名又は名称】　株式会社リコー

【代理人】

【識別番号】　100104190

【弁理士】

【氏名又は名称】　酒井　昭徳

【手数料の表示】

【予納台帳番号】　041759

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9810808

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【特許請求の範囲】

【請求項１】　文書の内容に基づいて文書の分類をおこなう文書分類装置において、

文書データを入力する入力手段と、

前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、

前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、

前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、

前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、

前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、

を備えたことを特徴とする文書分類装置。

【請求項２】　文書の内容に基づいて文書の分類をおこなう文書分類装置において、

文書データを入力する入力手段と、

前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、

前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、

前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、

前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、

出証特平１１－３０７２８０８

前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示する表示手段と、

前記表示手段に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するクラスタ選択指示手段と、

前記クラスタ選択指示手段により選択されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、

を備えたことを特徴とする文書分類装置。

【請求項３】　前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトルを前記分類手段により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するベクトル修正手段と、

を備え、

前記分類手段は、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする請求項２に記載の文書分類装置。

【請求項４】　前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正する文書表現空間修正手段と、

を備え、

前記分類手段は、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項２に記載の文書分類装置。

【請求項５】　前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正する文書表現空間修正手段と、

を備え、

前記分類手段は、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項３に記載の文書分類装置。

【請求項６】　前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与手段を備え、

前記表示手段は、前記クラスタ特徴を表示するとともに、前記選択情報付与手段により付与された選択情報を表示することを特徴とする請求項２または４に記載の文書分類装置。

【請求項７】　前記分類体系記憶手段は、前記選択指示手段により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報を分類体系の構成要素として記憶することを特徴とする請求項２～６に記載の文書分類装置。

【請求項８】　文書の内容に基づいて文書の分類をおこなう文書分類方法において、

文書データを入力する入力工程と、

前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、

前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、

前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、

前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、

前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、

を含んだことを特徴とする文書分類方法。

【請求項９】　文書の内容に基づいて文書の分類をおこなう文書分類方法において、

文書データを入力する入力工程と、

前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、

前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、

前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、

前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、

前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示する表示工程と、

前記表示工程に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するクラスタ選択指示工程と、

前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、

を含んだことを特徴とする文書分類方法。

【請求項１０】　前記ベクトル生成工程により生成された文書特徴ベクトルを前記分類工程により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するベクトル修正工程と、

を含み、

前記分類工程は、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする請求項９に記載の文書分類方法。

【請求項１１】　前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正する文書表現空間修正工程と、

を含み、

前記分類工程は、前記文書表現空間修正工程により修正された文書表現空間を

用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項９に記載の文書分類方法。

【請求項１２】　前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正する文書表現空間修正工程と、

を含み、

前記分類工程は、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項１０に記載の文書分類方法。

【請求項１３】　前記分類工程により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与工程を含み、

前記表示工程は、前記クラスタ特徴を表示するとともに、前記選択情報付与工程により付与された選択情報を表示することを特徴とする請求項９または１１に記載の文書分類方法。

【請求項１４】　前記分類体系生成工程は、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成することを特徴とする請求項９～１３に記載の文書分類方法。

【請求項１５】　前記請求項８～１４のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

この発明は、入力された複数の文書をその文書の内容に基づいて分類をおこなう、特に文書分類の際に算出される分類カテゴリー（体系）を精錬化する文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを

記録したコンピュータ読み取り可能な記録媒体に関する。

【０００２】

【従来の技術】

近年、インターネット等のネットワーク技術の普及により国内外の大量の電子化文書へのアクセスが可能になり、それに比例して業務上電子的に蓄積される情報の量も飛躍的に拡大した。その中で収集した大量の文書情報を意味あるカテゴリー（体系）に分類する等の知的作業の必要性が高まってきている。

【０００３】

これらの大量の文書情報を意味的に分類するという作業の目的は、以下のようなものである。まず第１に、検索容易性の向上が考えられる。これは、膨大な文書群を分類名称（内容名）を手がかりに検索できるので検索が比較的容易である。

【０００４】

第２に、情報群全体の把握が考えられる。これは、文書群全体がどのような内容（個々の分類）で構成されているかを把握する。しかし、大量の文書情報を操作者が手動で分類する場合、正確な分類をすることはできるが、分類に係る人的・時間的コストが膨大なものになるため、近年の文書分類の蓄積量の膨大さから、文書情報の自動分類装置が提案されるようになってきた。

【０００５】

文書自動分類装置の従来技術としては、たとえば、特開平７－３６８９７号公報に記載されているように、文書を、単語を特徴とする文書ベクトルとみなし、クラスタリング手法を用いてこれらの文書ベクトルを群分けし、群分けした文書ベクトルに基づいて文書の自動分類をおこなうものがある。

【０００６】

また、「Ｐｒｏｊｅｃｔｉｏｎｓ ｆｏｒ Ｅｆｆｉｃｉｅｎｔ Ｄｏｃｕｍｅｎｔ Ｃｌｕｓｔｅｒｉｎｇ（著者名：Ｈｉｎｒｉｃｈ Ｓｃｈｕｔｚｅ ａｎｄ Ｃｒａｉｎｇ Ｓｉｌｖｅｒｓｔｅｉｎ，　学会名：ＡＣＭ，　論文名：Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ＳＩＧＩＲ，　ページ：７４－８１，　発行年：１９９７）」においては、潜在的意味空間において文書分類を実施しているも

のがある。そのほかの方法としては、確率論的アプローチを用いる方法等が考えられる。

【０００７】

【発明が解決しようとする課題】

しかしながら、上記従来技術の文書分類装置は、本質的には単語で構成される多次元空間に布置した文書を統計的な分類をする方法であるため、分類結果は単語のいわゆる振る舞いという観点から統計的に求められたものにすぎず，分類の結果、算出される各クラスタ（分類された個々の文書の部分集合）が操作者（利用者）に理解不能な場合がある。

【０００８】

また、どのような分類結果が得られるかは分類対象のクラスタそのものに依存するため、最適な分類結果について定義することが困難であるという問題点があった。特に、上記情報群全体の把握に関し、多様な操作者の意図により要求される分類も異なるため、一度の分類作業で、操作者の所望する結果を得ることが困難であるという問題点があった。

【０００９】

このように、文書分類の結果は、多くのいわゆるノイズを含んだものであると解釈することができ、その一部についてのみが操作者にとって有益な場合が多いという問題点があった。

【００１０】

この発明は、上述した従来例による問題点を解消するため、任意の文書集合にどのような内容が含まれるかを漸次的に収集することができる文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することを目的とする。

【００１１】

【課題を解決するための手段】

上述した課題を解決し、目的を達成するため、請求項１の発明に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、文書データを入力する入力手段と、前記入力手段により入力された文書データを解

析して言語解析情報を得る言語解析手段と、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、を備えたことを特徴とする。

【００１２】

この請求項１の発明によれば、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

【００１３】

また、請求項２の発明に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、文書データを入力する入力手段と、前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示する表示手段と、前記表示手段に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するクラスタ選択指示手段と、前記クラスタ選択指示手段により選択されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、を備えたことを特徴とする。

【００１４】

この請求項２の発明によれば、選択されたクラスタのみを用いており、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【００１５】

また、請求項３の発明に係る文書分類装置は、請求項２の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトルを前記分類手段により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するベクトル修正手段と、を備え、前記分類手段が、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする。

【００１６】

この請求項３の発明によれば、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【００１７】

また、請求項４の発明に係る文書分類装置は、請求項２の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正する文書表現空間修正手段と、を備え、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【００１８】

この請求項４の発明によれば、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【００１９】

また、請求項５の発明に係る文書分類装置は、請求項３の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選

択されたクラスタ特徴に基づいて修正する文書表現空間修正手段と、を備え、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【００２０】

この請求項５の発明によれば、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【００２１】

また、請求項６の発明に係る文書分類装置は、請求項２または４の発明において、前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与手段を備え、前記表示手段が、前記クラスタ特徴を表示するとともに、前記選択情報付与手段により付与された選択情報を表示することを特徴とする。

【００２２】

この請求項６の発明によれば、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができる。

【００２３】

また、請求項７の発明に係る文書分類装置は、請求項２～６の発明において、前記分類体系記憶手段が、前記選択指示手段により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報を分類体系の構成要素として記憶することを特徴とする。

【００２４】

この請求項７の発明によれば、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できることので、分類体系の利用価値を向上させることができる。

【００２５】

また、請求項８の発明に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、文書データを入力する入力工程と、前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、を含んだことを特徴とする。

【００２６】

この請求項８の発明によれば、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

【００２７】

また、請求項９の発明に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、文書データを入力する入力工程と、前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示する表示工程と、前記表示工程に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するクラスタ選択指示工程と、前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、を含んだことを特徴とする。

【００２８】

この請求項９の発明によれば、選択されたクラスタのみを用いており、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【００２９】

また、請求項１０の発明に係る文書分類方法は、請求項９の発明において、前記ベクトル生成工程により生成された文書特徴ベクトルを前記分類工程により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するベクトル修正工程と、を含み、前記分類工程が、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする。

【００３０】

この請求項１０の発明によれば、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【００３１】

また、請求項１１の発明に係る文書分類方法は、請求項９の発明において、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正する文書表現空間修正工程と、を含み、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【００３２】

この請求項１１の発明によれば、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【００３３】

また、請求項１２の発明に係る文書分類方法は、請求項１０の発明において、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正する文書表現空間修正工程と、を含み、前記分類工程が、前記文書

表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【００３４】

この請求項１２の発明によれば、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【００３５】

また、請求項１３の発明に係る文書分類方法は、請求項９または１１の発明において、前記分類工程により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与工程を含み、前記表示工程が、前記クラスタ特徴を表示するとともに、前記選択情報付与工程により付与された選択情報を表示することを特徴とする。

【００３６】

この請求項１３の発明によれば、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができる。

【００３７】

また、請求項１４の発明に係る文書分類方法は、請求項９～１３の発明において、前記分類体系生成工程が、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成することを特徴とする。

【００３８】

この請求項１４の発明によれば、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できることので、分類体系の利用価値を向上させることができる。

【００３９】

また、請求項１５の発明に係る記憶媒体は、請求項８～１４に記載された方法

をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項８～１４の動作をコンピュータによって実現することが可能である。

【００４０】

【発明の実施の形態】

以下に添付図面を参照して、この発明に係る文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体の好適な実施の形態を詳細に説明する。

【００４１】

なお、以下説明する実施の形態においては、上記のように多くのノイズを含んだものであるとの解釈に基づいて、一回の文書集合からの話題（内容）抽出と位置づけ、文書分類のためのパラメータ（対象文書集合やクラスタ数、類似度測度、ストップワード等）を変化させながら複数化の分類を実行させ、その結果を保持・統合する手段を設けることで、任意の文書集合にどのような内容が含まれるかを漸次的に収集するものである。

【００４２】

（実施の形態１）

まず、この発明の実施の形態１による文書分類装置を構成する情報処理システム全体のハードウエア構成を説明する。図１は、本実施の形態による文書分類装置を構成する情報処理システム全体のハードウエア構成を示す説明図である。

【００４３】

図１において、実施の形態１による文書分類装置を構成する情報処理システムは、サーバー／クライアント方式で構成されている。すなわち、サーバー１０１と複数のクライアント１０２がネットワーク１０３によって接続されている。クライアント１０２は、分類データ等の加工データの生成、サーバー１０１への指示、分類結果等の加工処理結果の表示などをおこなう。一方、クライアント１０２からの指示にしたがって、サーバー１０１は文書（テキスト）分類等の加工処理を膨大な数値演算によりおこない、その処理の結果をクライアント１０２へ送る。

【００４４】

分類処理の場合、より具体的には、サーバー１０１においては、テキスト分類処理（前処理、クラスタリング処理）がおこなわれ、クライアント１０２においては、分類データ生成、処理実行指示、テキスト分類結果表示等がおこなわれる。サーバー１０１における処理は、上述のように、「前処理」と「分類処理」の二つに分かれており、その処理はデータによっては非常に負荷が大きくなる。したがって、サーバー１０１は「前処理」と「分類処理」がそれぞれ一つずつしか処理をおこなわないようにマネージャプロセスが処理受付リストを作成して管理する。

【００４５】

また、サーバー１０１とクライアント１０２との間のデータのやりとりはファイル共有という方法を用いる。すなわち、分類処理等の加工処理に用いるファイルをサーバー１０１上の共有フォルダに作成することにより両者はデータのやりとりをおこなう。したがって、クライアント１０２からはサーバー１０１の共有フォルダをネットワーク共有して利用することが可能である。

【００４６】

つぎに、サーバー１０１およびクライアント１０２のハードウエア構成について説明する。図２は、実施の形態１による文書分類装置を構成する情報処理システムにおけるサーバー１０１をハードウエア的に示す説明図である。サーバー１０１は、たとえばワークステーション（ＷＳ）等が用いられる。

【００４７】

図２において、２０１はサーバー１０１全体を制御するＣＰＵを、２０２はブートプログラム等を記憶したＲＯＭを、２０３はＣＰＵ２０１のワークエリアとして使用されるＲＡＭ２０３を、２０４は通信回線２０５を介してネットワーク１０３に接続され、そのネットワーク１０３と内部のインターフェイスを司るインターフェイス（Ｉ／Ｆ）を、２０６はデータを記憶するディスク装置を示している。２００は上記各部を結合させるためのバスを示している。

【００４８】

そのほか、文書情報、画像情報、機能情報等を表示するディスプレイ２０８や

，データを入力するためのキーボード２０９およびマウス２１０等が同様に接続されていてもよい。さらに、ディスク装置２０６には、クライアント１０２との間のデータのやりとりをするための共有フォルダ２０７が設けられている。

【００４９】

また、図３は、実施の形態１による文書分類装置を構成する情報処理システムにおけるクライアント１０２をハードウエア的に示す説明図である。クライアント１０２は、たとえばパーソナルコンピュータ（ＰＣ）等が用いられる。

【００５０】

図３において、３０１はシステム全体を制御するＣＰＵを、３０２はブートプログラム等を記憶したＲＯＭを、３０３はＣＰＵ３０１のワークエリアとして使用されるＲＡＭを、３０４はＣＰＵ３０１の制御にしたがってＨＤ（ハードディスク）３０５に対するデータのリード／ライトを制御するＨＤＤ（ハードディスクドライブ）を、３０５はＨＤＤ３０４の制御で書き込まれたデータを記憶するＨＤを、３０６はＣＰＵ３０１の制御にしたがってＦＤ（フロッピーディスク）３０７に対するデータのリード／ライトを制御するＦＤＤ（フロッピーディスクドライブ）を、３０７はＦＤＤ３０６の制御で書き込まれたデータを記憶する着脱自在のＦＤを、３０８はドキュメント、画像、機能情報等を表示するディスプレイをそれぞれ示している。

【００５１】

また、３０９は通信回線３１０を介してネットワーク１０３に接続され、そのネットワーク１０３と内部のインターフェイスを司るインターフェイス（Ｉ／Ｆ）を、３１１は文字、数値、各種指示等の入力のためのキーを備えたキーボードを、３１２はカーソルの移動や範囲選択、あるいは表示画面に表示されたアイコンやボタンの押下やウインドウの移動やサイズの変更等をおこなうマウスを、３１３はＯＣＲ（Ｏｐｔｉｃａｌ　Ｃｈａｒａｃｔｅｒ　Ｒｅａｄｅｒ）機能を備えた画像を光学的に読み取るスキャナを、３１４は分類結果を含むデータの内容等を印刷するプリンタを、３１５は上記各部を結合するためのバスをそれぞれ示している。また、ＨＤ３０５にはワープロソフト、表計算ソフト等のアプリケーションソフト３１６が記憶されている。

【００５２】

つぎに、実施の形態１による文書分類装置の機能的構成について説明する。図４は、実施の形態１による文書分類装置の構成を機能的に示すブロック図である。

【００５３】

図４のブロック図において、文書分類装置は、入力部４０１と、言語解析部４０２と、ベクトル生成部４０３と、分類部４０４と、分類パラメータ指示部４０５と、分類結果記憶部４０６と、クラスタ特徴表示部４０７と、クラスタ特徴算出部４０８と、分類体系記憶部４０９と、クラスタ選択指示部４１０と、分類体系閲覧操作部４１１と、を含む構成である。

【００５４】

入力部４０１、言語解析部４０２、ベクトル生成部４０３、分類部４０４、分類パラメータ指示部４０５、分類結果記憶部４０６、クラスタ特徴表示部４０７、クラスタ特徴算出部４０８、分類体系記憶部４０９、クラスタ選択指示部４１０、分類体系閲覧操作部４１１は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【００５５】

ここで、入力部４０１は、文書データを入力するものであり、たとえば、キーボード２０９または３１１、スキャナ３１３、ＯＣＲ機能を備えたスキャナ３１３、またはネットワーク１０３を経由して文書や文書群を得ることができるＩ／Ｆ２０４または３０９等である。

【００５６】

また、入力部４０１は、上記以外に、文書データを取得することができるものであれば、それらのすべてを含む。たとえば、文書データがデータベース化されている場合に、そのデータベースが記録された媒体を本実施の形態の文書分類装置に組み入れた場合も文書データの入力とする。

【００５７】

また、言語解析部４０２は、入力部４０１により入力された文書データを解析して言語解析情報を得るものであり、ベクトル生成部４０３は、言語解析部４０２により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するものである。

【００５８】

また、分類部４０４は、ベクトル生成部４０３により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成するものであり、分類パラメータ指示部４０５は、分類パラメータを指示するものであり、たとえば、キーボード２０９または３１１、マウス２１０または３１２、またはネットワーク１０３を経由して指示情報を得ることができるＩ／Ｆ２０４または３０９等である。

【００５９】

また、分類結果記憶部４０６は、分類部４０４により分類された結果、すなわち、分類された文書の部分集合に関する情報を記憶するものである。また、クラスタ特徴表示部４０７は、クラスタ特徴算出部４０８により算出されたクラスタ特徴を表示する。

【００６０】

クラスタ特徴算出部４０８は、分類４０４により生成された文書の部分集合の特徴であるクラスタ特徴を算出するものである。また、分類体系記憶部４０９は、クラスタ選択指示部４１０により選択されたクラスタ特徴を分類体系の構成要素として記憶するものである。

【００６１】

クラスタ選択指示部４１０は、表示部４０７に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するものである。また、分類体系閲覧操作部４１１は、分類体系記憶部４０９に記憶されたデータを閲覧したい場合に、その閲覧の操作をおこなうものである。

【００６２】

つぎに、文書集合に含まれる話題（内容）を抽出することが重要となる好適な例を、アンケート調査等により得られた自由記述回答の分析場面を想定し、その

具体例を用いて説明する。

【００６３】

近年、たとえば、インターネット等を介して短期間に数千～数万件の自由記述回答を回収することが可能であり、このような機能を用いて大量のテキスト情報の収集をおこなうことができる。

【００６４】

アンケート調査により得られた大量のテキスト情報の収集の例として、「オフィスのネットワーク化による無駄を挙げてください」という質問に対して文書で答えた一つの回答記述を文書とすると、文書集合（クラスタ）は１件ごとの回答の集合ということになる。

【００６５】

ここで、操作者（アンケートの分析者）は、そのニーズの一つとして、意見集合（文書集合）にどのような種類の意見（話題）が含まれており、意見の概略を把握したい場合がある。このようなニーズを満たすべく、話題の抽出を類似する意見のまとまり（分類）により実現し、アンケート結果にどのような種類の意見が含まれているかを抽出する。

【００６６】

文書分類は、典型的には大きく分けてつぎの３段階のステップから構成される。第１ステップでは、入力部４０１により入力された各文書（意見）について、言語解析部４０２が、各文書に含まれる単語（あるいは、特定の連続する文字列）を抽出する。この際、たとえば、形態素形跡等の言語解析アルゴリズムが用いられる。

【００６７】

第２ステップでは、抽出された単語を列とし、各文書を行とし、要素を単語の出現頻度とした「単語」×「文書」の行列が生成される。なお、一般的な形態素解析機能と構文解析機能を有する言語解析ツールを用いると単語抽出のほかに、単語の品詞情報、複合語（フレーズ）、構文情報等の同時に取得することができ、こうした情報を上記単語×文書の行列を生成する際、考慮することができる。

【００６８】

ベクトル生成部４０３は、この「単語」×「文書」の行列に基づいて単語で構成される多次元空間内に各文書をベクトル表現する。これには、以下の方法があり、本実施の形態においては、すべての方法を実装している。

【００６９】

（１）行列の列成分をそのまま利用する方法、

（２）各文書の長さ（文字の数やページ数等）や分類対象全体の文書集合内での各単語の出現頻度を考慮して値の重み付けをする方法、

（３）上記行列から文書間の内積行列を算出し、これに特異値分解（たとえば、因子分析や主成分分析、数量化理論第３類等を利用しておこなわれる）を適用して潜在的意味空間を構成する方法、

等である。

【００７０】

また、「Ｒｅｐｒｅｓｅｎｔａｔｉｎｇ Ｄｏｃｕｍｅｎｔｓ Ｕｓｉｎｇ ａｎ Ｅｘｐｌｉｃｉｔ Ｍｏｄｅｌ ｏｆ Ｔｈｅｉｒ Ｓｉｍｉｌａｒｉｔｉｅｓ（著者名：Ｂｒｉａｎ Ｔ. Ｂａｒｔｅｌｌ，Ｇａｒｒｉｓｏｎ Ｗ. Ｃｏｔｔｒｅｌｌ，ａｎｄ Ｒｉｃｈａｒｄ Ｋ. Ｂｅｌｅｗ，論文名：Ｊｏｕｒｎａｌ ｏｆ ｔｈｅ Ａｍｅｒｉｃａｎ Ｓｏｃｉｅｔｙ ｆｏｒ Ｉｎｆｏｒｍａｔｉｏｎ Ｓｃｉｅｎｃｅ，学会名：ｔｈｅ Ａｍｅｒｉｃａｎ Ｓｏｃｉｅｔｙ ｆｏｒ Ｉｎｆｏｒｍａｔｉｏｎ Ｓｃｉｅｎｃｅ，ページ：２５４－２７１，Ｖｏｌ. ４６ Ｎｏ. ４，発行年：１９９５）」においては、上記潜在的意味空間への変換手法を一般化し、文書間の内積行列に、文書が有するほかの文書への参照情報から生成される共参照情報などを付加した行列を用いて、これらの類似性を反映する空間へ文書や単語を射影するための表現空間変換関数を導出しているものもあり、この方法も利用することができる。

【００７１】

第３ステップでは、分類部４０４が、文書特徴ベクトルの類似度を用いて文書を分類する。具体的には分類対象データに対してカイ自乗法の手法、判別分析の方法、クラスタリングの方法等を適用することにより分類が実行される。

【００７２】

また、類似度としては、内積や余弦、ユークリッド距離、マハラノビスの距離等が考えられ、本実施の形態においては、いずれの方法を用いてもよい。

【００７３】

また、クラスタリングのアルゴリズムに関しても様々なものが公知になっている。クラスタリングは大別して階層型クラスタリングと非階層型クラスタリングが考えられるが、本実施の形態においては、いずれの方法を用いてもよい。

【００７４】

また、分類パラメータ指示部４０５は、分類部４０４が文書特徴ベクトルを分類するための分類パラメータを指示する。分類部４０４は、分類パラメータ指示部４０５により指示された分類パラメータにしたがって内部に保持される文書特徴ベクトルを分類する。

【００７５】

このようにして、第１ステップ～第３ステップの各処理を実行することにより第１回目の文書分類が終了すると、分類結果は分類結果記憶部４０６により保持される。

【００７６】

引き続き、クラスタ特徴算出部４０８が、分類結果がどのようなクラスタを得ることができたのかを示す特徴、すなわちクラス特徴を算出する。典型的には各クラスタに所属する文書、あるいはその文書の一部を算出するが、その際、クラスタの重心との類似度に基づいて文書をソーティングして出力する。

【００７７】

そのほか、クラスタ内で最頻の単語、クラスタに所属する文書数、クラスタ内での文書のばらつきの程度をあらわすクラスタ内の標準偏差のような数値をクラスタの特徴を表現するものとして算出する。

【００７８】

これらのクラスタの特徴情報は、操作者に対して出力（表示）されたクラスタがどのようなもの（どのような特徴を有するもの）かを把握させるために算出されるものであり、操作者に対してクラスタの特徴を示すものであれば、上記の内

容（特徴）以外のものであってもよい。

【００７９】

また、クラスタ特徴算出部４０８は、上記のようにクラスタの特徴を示すもの以外に、クラスタ間の関係を示す情報も算出する。階層型クラスタリングの場合は、その上位あるいは下位のクラスタを、非階層型クラスタリングの場合は、クラスタ重心間の類似度に基づく近接のクラスタを算出する。

【００８０】

つぎに、クラスタ特徴表示部４０７によるクラスタ特徴の表示およびクラスタ選択の内容について説明する。図５は、実施の形態１による文書分類装置のクラスタ特徴表示部４０７の表示の一例を示す説明図である。

【００８１】

図５において、クラスタ単位で操作者ができるようになっており、各クラスタは「クラスタＩＤ」欄５０１、「メンバー数」欄５０２、「頻度の高い単語」欄５０３、「文書内容」欄５０４、「重心との類似度」欄５０５等の項目から構成される。

【００８２】

「クラスタＩＤ」欄５０１には、クラスタのＩＤを示す番号が通し番号で付与され、表示される。「メンバー数」欄５０２はクラスタに所属する文書あるいは文書の一部の数が算出され、表示される。その中で頻度の高い単語が抽出され「頻度の高い単語」欄５０３に表示される。「文書内容」欄５０４には文書の内容が表示され、「重心との類似度」欄５０５には、数値化された重心との類似度が表示される。これにより、操作者の理解容易性が向上する。

【００８３】

操作者は、表示された情報（特徴量）に基づいてクラスタについてその特徴を把握することができる。ここで、内容（特徴）が理解可能なクラスタが一つでもあれば、クラスタ選択指示部４１０によりクラスタを選択することができる。

【００８４】

より具体的には、マウス２１０または３１２等によって、表示されているクラスタの所定の位置、たとえば、「クラスタＩＤ」欄５０１へカーソル５１０を移

動させ、その位置でクリックすることにより、当該クラスタＩＤのクラスタ全体を選択することができる。

【００８５】

図５においては、「クラスタＩＤ」欄５０１がクリックされ、これにより、クラスタ全体が反転表示しており、当該クラスタ（クラスタＩＤ「１」）が選択されたことを示している。

【００８６】

また、操作者は、内容が理解可能であるクラスタが存在しない場合は、分類パラメータ指示部４０５により分類パラメータの再設定をおこない、再度分類実行をおこなうことができる。

【００８７】

クラスタ選択指示部４１０により選択されたクラスタＩＤに関するデータは分類体系記憶部４０９へ送信される。分類体系記憶部４０９は、このクラスタＩＤに関するデータに基づいてクラスタ特徴算出部４０８からクラスタに関する上記特徴量を検索し記憶する。

【００８８】

また、分類体系記憶部４０９は、同様に、分類結果記憶部４０６から分類結果を検索し記憶する。さらに、分類体系記憶部４０９は、操作者により入力されたクラスタに関するコメント（たとえば、「ネットワークの維持費が高い」等）の情報を併せて記憶することもできる。このように、操作者が作成した情報を分類体系の構成要素として記憶することにより、分類体系の利用価値がより向上する。

【００８９】

なお、分類体系記憶部４０９により記憶されたデータは、別途閲覧操作用のインターフェイスを設けることにより、選択・保持するクラスタの内容の閲覧や、クラスタ間の意味的な関連を手動であるいは、保持されているクラスタ重心間の類似度等を用いて自動で、構造化・体系化することができる。

【００９０】

つぎに、実施の形態１の文書分類装置の一連の処理の手順について説明する。

図６は、実施の形態１による文書分類装置の一連の処理の手順を示すフローチャートである。図６のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップＳ６０１）。

【００９１】

つぎに、入力された文書の言語が解析され（ステップＳ６０２）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成される（ステップＳ６０３）。

【００９２】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップＳ６０４肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップＳ６０５）、その結果、すなわち、クラスタに関する情報を記憶する（ステップＳ６０６）。

【００９３】

つぎに、分類されたクラスタの特徴を算出し（ステップＳ６０７）、算出された結果を表示する（ステップＳ６０８）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップＳ６０９）、選択されなかった場合（ステップＳ６０９否定）は、ステップＳ６０４へ移行し、再度分類パラメータの指示があるのを待つ（ステップＳ６０４）。

【００９４】

一方、ステップＳ６０９において、クラスタが選択された場合（ステップＳ６０９肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップ６１０）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。これにより、一連の処理を終了する。

【００９５】

以上説明したように、実施の形態１による文書分類装置によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することができる。

【００９６】

したがって、分類部４０４によりクラスタを得ることができるとともに、クラスタ特徴算出部４０８・分類体系記憶部４０９により、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

【００９７】

また、クラスタ選択指示部４１０により選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【００９８】

（実施の形態２）

さて、上述した実施の形態１に加えて、以下に説明する実施の形態２のように、さらにベクトル記憶部と、ベクトル修正部とを含む構成とするようにしてもよい。

【００９９】

実施の形態２による文書分類装置を構成する情報処理システムは、図１に示したように実施の形態１と同様であるので、その説明は省略する。また、サーバー１０１およびクライアント１０２のハードウエア構成についても、図２・図３に示したように実施の形態１と同様であるので、その説明は省略する。

【０１００】

つぎに、実施の形態２による文書分類装置の機能的構成について説明する。図７は、この発明の実施の形態２による文書分類装置の構成を機能的に示すブロック図である。図７において、実施の形態１の図４と同一のものに関しては同じ符号を付して、その説明を省略する。

【０１０１】

図７のブロック図において、文書分類装置は、入力部４０１、言語解析部４０２、ベクトル生成部４０３、分類部４０４、分類パラメータ指示部４０５、分類結果記憶部４０６、クラスタ特徴表示部４０７、クラスタ特徴算出部４０８、分類体系記憶部４０９、クラスタ選択指示部４１０、分類体系閲覧操作部４１１の

ほか、ベクトル記憶部７０１と、ベクトル修正部７０２とを含む構成である。

【０１０２】

ベクトル記憶部７０１は、ベクトル生成部４０３により生成された文書特徴ベクトルを記憶するものである。また、ベクトル修正部７０２は、文書特徴ベクトル記憶部７０１により記憶された文書特徴ベクトルを分類部４０４により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するものである。

【０１０３】

また、分類部４０４は、ベクトル修正部７０２により修正された文書特徴ベクトルに基づいて文書を分類する。

【０１０４】

なお、ベクトル記憶部７０１、ベクトル修正部７０２は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０１０５】

ベクトル生成部４０３において生成された文書特徴ベクトル（列ベクトル）・単語（単語特徴）ベクトル（行ベクトル）はベクトル記憶部７０１によって記憶される。これは、次回以降の分類実行の際に利用する文書特徴ベクトルを確保するためである。

【０１０６】

ベクトル修正部７０２は、クラスタ選択指示部４１０により選択されたクラスタに所属する文書のすべてあるいはその一部の文書を除き、次回以降もこれらの文書が除かれるよう削除する。削除された文書特徴ベクトルはベクトル記憶部７０１により記憶される。

【０１０７】

この結果、ベクトル記憶部７０１に記憶されているベクトルデータのうち、選択されたクラスタに所属する文書（もしくは操作者に指定されたその一部）列ベ

クトルを除いたものが次回以降の分類実行の際に利用されるデータとなる。

【０１０８】

つぎに、実施の形態２の文書分類装置の一連の処理の手順について説明する。図８は、実施の形態２よる文書分類装置の一連の処理の手順を示すフローチャートである。図２のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップＳ８０１）。

【０１０９】

つぎに、入力された文書の言語が解析され（ステップＳ８０２）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップＳ８０３）、生成された文書特徴ベクトルが記憶される（ステップＳ８０４）。

【０１１０】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップＳ８０５肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップＳ８０６）、その結果、すなわち、クラスタに関する情報を記憶する（ステップＳ８０７）。

【０１１１】

つぎに、分類されたクラスタの特徴を算出し（ステップＳ８０８）、算出された結果を表示する（ステップＳ８０９）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップＳ８１０）、選択されなかった場合（ステップＳ８１０否定）は、ステップＳ８０５へ移行し、再度分類パラメータの指示があるのを待つ（ステップＳ８０５）。

【０１１２】

一方、ステップＳ８１０において、クラスタが選択された場合（ステップＳ８１０肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップ８１１）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップＳ８１２）。

【０１１３】

ステップＳ８１２において、繰り返して処理をおこなう旨の指示があった場合（ステップＳ８１２肯定）は、選択されたクラスタに所属する文書のすべてあるいはその一部の文書を除くように文書特徴ベクトルを修正する（ステップＳ８１３）。その後、ステップＳ８０５へ移行し、以後、ステップＳ８０５～Ｓ８１３の各処理を繰り返しおこなう。

【０１１４】

一方、ステップＳ８１２において、繰り返して処理をおこなう旨の指示がない場合（ステップＳ８１２否定）は、これにより、一連の処理を終了する。

【０１１５】

以上説明したように、実施の形態２による文書分類装置によれば、ベクトル修正部７０１により、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【０１１６】

（実施の形態３）

さて、上述した実施の形態２においては、ベクトル記憶部およびベクトル修正部とを含む構成であったが、以下に説明する実施の形態３のように、ベクトル修正部に代わりに、文書表現空間修正部を含む構成とするようにしてもよい。

【０１１７】

実施の形態３による文書分類装置を構成する情報処理システムは、図１に示したように実施の形態１と同様であるので、その説明は省略する。また、サーバー１０１およびクライアント１０２のハードウエア構成についても、図２・図３に示したように実施の形態１と同様であるので、その説明は省略する。

【０１１８】

つぎに、実施の形態３による文書分類装置の機能的構成について説明する。図９は、この発明の実施の形態３による文書分類装置の構成を機能的に示すブロック図である。図９において、実施の形態１の図４と同一のものに関しては同じ符号を付して、その説明を省略する。

【０１１９】

図９のブロック図において、文書分類装置は、入力部４０１、言語解析部４０

２、ベクトル生成部４０３、分類部４０４、分類パラメータ指示部４０５、分類結果記憶部４０６、クラスタ特徴表示部４０７、クラスタ特徴算出部４０８、分類体系記憶部４０９、クラスタ選択指示部４１０、分類体系閲覧操作部４１１のほか、ベクトル記憶部９０１と、文書表現空間修正部９０２とを含む構成である。

【０１２０】

ベクトル記憶部９０１は、ベクトル生成部４０３により生成された文書特徴ベクトルを記憶するものである。また、文書表現空間修正部９０２は、文書特徴ベクトル記憶部９０１により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示部４１０により選択されたクラスタ特徴に基づいて修正するものである。

【０１２１】

また、分類部４０４は、文書表現空間修正部９０２により修正された文書表現空間を用いて、ベクトル生成部４０３により生成された文書特徴ベクトル間の類似度に基づいて文書を分類する。

【０１２２】

なお、ベクトル記憶部９０１、文書表現空間修正部９０２は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０１２３】

つぎに、文書表現空間修正部９０２の内容について説明する。実施の形態２におけるベクトル修正部７０２にあっては、既知になったクラスタの影響を排除するために文書特徴ベクトルを除去するが、文書特徴ベクトルを表現する多次元空間自体の変更はおこなわれない。

【０１２４】

したがって、前回の分類実行の結果、操作者により選択されたクラスタの形成特徴を次回の分類実行の際に排除したい場合は、文書ベクトルを表現する空間自

体の変更が必要となる。

【０１２５】

そこで、文書表現空間修正部９０２を備え、文書表現空間の修正をおこなうものである。ここで、文書表現空間の特徴次元を変更する例として、操作者により選択されたクラスタの重心と類似度の高い特徴次元の削除をおこなうことについて説明する。

【０１２６】

操作者により選択されたクラスタの重心はベクトルとして表現することができるので、このクラスタ重心ベクトルとベクトル記憶部９０１に記憶されている文書表現空間の各特徴次元との類似度を算出することにより、類似度の高い特徴次元を判別する。

【０１２７】

なお、類似の測度としては、余弦、内積、ユークリッド距離、マハラノビス距離等を用いる。また、判別に関してはある類似度以上を削除対象として採用するようなしきい値処理による判別や、類似度の高い順にある一定数を削除対象として採用する定数処理による判別を用いる。また、判別分析等も用いることができる。

【０１２８】

文書表現空間修正部９０２は、上述のような削除対象の特徴次元を算出して、特徴次元の削除をおこなう。特徴次元の削除は、ベクトル記憶部９０１に記憶されている「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することによりおこなう。文書表現空間修正部９０２により修正された文書ベクトルは、次回以降の分類のために、ベクトル記憶部９０１に記憶される。

【０１２９】

つぎに、実施の形態３の文書分類装置の一連の処理の手順について説明する。図１０は、実施の形態３よる文書分類装置の一連の処理の手順を示すフローチャートである。図１０のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップＳ１００１）。

【０１３０】

つぎに、入力された文書の言語が解析され（ステップＳ１００２）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップＳ１００３）、生成された文書特徴ベクトルが記憶される（ステップＳ１００４）。

【０１３１】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップＳ１００５肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップＳ１００６）、その結果、すなわち、クラスタに関する情報を記憶する（ステップＳ１００７）。

【０１３２】

つぎに、分類されたクラスタの特徴を算出し（ステップＳ１００８）、算出された結果を表示する（ステップＳ１００９）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップＳ１０１０）、選択されなかった場合（ステップＳ１０１０否定）は、ステップＳ１００５へ移行し、再度分類パラメータの指示があるのを待つ（ステップＳ１００５）。

【０１３３】

一方、ステップＳ１０１０において、クラスタが選択された場合（ステップＳ１０１０肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップ１０１１）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップＳ１０１２）。

【０１３４】

ステップＳ１０１２において、繰り返して処理をおこなう旨の指示があった場合（ステップＳ１０１２肯定）は、「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することにより文書表現空間を修正する（ステップＳ１０１３）。その後、ステップＳ１００５へ移行し、以後、ステップＳ１００５～Ｓ１０１３の各処理を繰り返しおこなう。

【０１３５】

一方、ステップＳ１０１２において、繰り返して処理をおこなう旨の指示がなかった場合（ステップＳ１０１２否定）は、これにより、一連の処理を終了する。

【０１３６】

以上説明したように、実施の形態３による文書分類装置によれば、前回の分類実行の結果、文書表現空間修正部９０２により操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【０１３７】

（実施の形態４）

さて、上述した実施の形態２または実施の形態３においては、ベクトル修正部または文書表現空間修正部のいずれか一方のみを含む構成であったが、以下に説明する実施の形態４のように、ベクトル修正部および文書表現空間修正部の両方を含む構成とするようにしてもよい。

【０１３８】

実施の形態４による文書分類装置を構成する情報処理システムは、図１に示したように実施の形態１と同様であるので、その説明は省略する。また、サーバー１０１およびクライアント１０２のハードウエア構成についても、図２・図３に示したように実施の形態１と同様であるので、その説明は省略する。

【０１３９】

つぎに、実施の形態４による文書分類装置の機能的構成について説明する。図１１は、この発明の実施の形態４による文書分類装置の構成を機能的に示すブロック図である。図１１において、実施の形態１の図４と同一のものに関しては同じ符号を付して、その説明を省略する。

【０１４０】

図１１のブロック図において、文書分類装置は、入力部４０１、言語解析部４０２、ベクトル生成部４０３、分類部４０４、分類パラメータ指示部４０５、分類結果記憶部４０６、クラスタ特徴表示部４０７、クラスタ特徴算出部４０８、分類体系記憶部４０９、クラスタ選択指示部４１０、分類体系閲覧操作部４１１

のほか、ベクトル記憶部１１０１と、ベクトル修正部１１０２と、文書表現空間修正部１１０３とを含む構成である。

【０１４１】

ベクトル記憶部１１０１は、ベクトル生成部４０３により生成された文書特徴ベクトルを記憶するものである。また、ベクトル修正部１１０２は、文書特徴ベクトル記憶部１１０１により記憶された文書特徴ベクトルを分類部４０４により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するものである。

【０１４２】

また、文書表現空間修正部１１０３は、文書特徴ベクトル記憶部１１０１により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示部４１０により選択されたクラスタ特徴に基づいて修正するものである。

【０１４３】

また、分類部４０４は、文書表現空間修正部１１０３により修正された文書表現空間を用いて、ベクトル修正部１１０２により修正された文書特徴ベクトル間の類似度に基づいて文書を分類する。

【０１４４】

なお、ベクトル記憶部１１０１、ベクトル修正部１１０２、文書表現空間修正部１１０３は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０１４５】

つぎに、ベクトル修正部１１０２および文書表現空間修正部１１０３の内容について説明する。実施の形態３においては、選択されたクラスタに所属する文書は次回以降の分類実行の際にも使用される。

【０１４６】

実施の形態４では、ベクトル修正部１１０２および文書表現空間修正部１１０

３の両方を具備することにより、選択されたクラスタに所属する文書を次回の分類実行の際に除去し、次回の分類実行の際には分類対象文書としないようにする。

【０１４７】

実施の形態３においては、話題抽出の側面を強調し、ある文書が複数の話題として分類される可能性を前提としており、たとえば、ネットワーク化に関する調査における「エンドユーザーがソフトウエアのインストール方法について聞いてくるのでシステム管理者としての仕事ができない」という回答について言えば、この意見は「ソフトウエアの操作方法理解に関する困難性」という話題として分類され得るし、「システム管理者の仕事の多忙さ」という話題で分類される可能性もある。

【０１４８】

実施の形態３においては、いずれにしても、「ソフトウエアの操作方法理解に関する困難性」というクラスタと「システム管理者の仕事の多忙さ」というクラスタの両方とも抽出したいというニーズに応えている。

【０１４９】

これとは反対に、操作者は、一度抽出した話題は既知であるので、次回の分類の際はなるべく異なる分類結果が欲しいとするケースも考えられる。実施の形態４では、このような要求に応えるため、ベクトル修正部１１０２により、ｎ回目の分類で選択されたクラスタに所属する文書のすべてまたはその一部を次回以降の分類を実行する際、分類対象から除去するものである。

【０１５０】

クラスタ選択指示部４１０により選択指示を受けたクラスタの所属文書はベクトル記憶部１１０１において列ベクトルの形式で記憶されているため、ベクトル修正部１１０２では劣ベクトルを除去することで、次回以降の分類実行用の分類対象文書集合を生成する。

【０１５１】

さらに、実施の形態３と同様に、選択されたクラスタにより文書表現空間修正部１１０３は、ベクトル記憶部１１０１に記憶されている行列から特徴次元を削

除する。

【０１５２】

つぎに、実施の形態４の文書分類装置の一連の処理の手順について説明する。図１２は、実施の形態４よる文書分類装置の一連の処理の手順を示すフローチャートである。図１２のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップＳ１２０１）。

【０１５３】

つぎに、入力された文書の言語が解析され（ステップＳ１２０２）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップＳ１２０３）、生成された文書特徴ベクトルが記憶される（ステップＳ１２０４）。

【０１５４】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップＳ１２０５肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップＳ１２０６）、その結果、すなわち、クラスタに関する情報を記憶する（ステップＳ１２０７）。

【０１５５】

つぎに、分類されたクラスタの特徴を算出し（ステップＳ１２０８）、算出された結果を表示する（ステップＳ１２０９）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップＳ１２１０）、選択されなかった場合（ステップＳ１２１０否定）は、ステップＳ１２０５へ移行し、再度分類パラメータの指示があるのを待つ（ステップＳ１２０５）。

【０１５６】

一方、ステップＳ１２１０において、クラスタが選択された場合（ステップＳ１２１０肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップ１２１１）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップＳ１２１２）。

【０１５７】

ステップＳ１２１２において、繰り返して処理をおこなう旨の指示があった場合（ステップＳ１２１２肯定）は、選択されたクラスタに所属する文書のすべてあるいはその一部の文書を除くように文書特徴ベクトルを修正する（ステップＳ１２１３）。

【０１５８】

ステップＳ１２１３に引き続き、「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することにより文書表現空間を修正する（ステップＳ１２１４）。その後、ステップＳ１２０５へ移行し、以後、ステップＳ１２０５～Ｓ１２１４を繰り返しおこなう。

【０１５９】

一方、ステップＳ１２１２において、繰り返して処理をおこなう旨に指示がない場合（ステップＳ１２１２否定）は、これにより、一連の処理を終了する。

【０１６０】

以上説明したように、実施の形態４よる文書分類装置によれば、ベクトル修正部１１０２が、既知になったクラスタの影響を排除し、かつ、文書表現空間修正部１１０３が、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【０１６１】

（実施の形態５）

さて、上述した実施の形態１または実施の形態３においては、繰り返し分類処理をおこなった場合に、ある文書が何度選択されたかその情報については考慮していなかったが以下に説明する実施の形態５のように、選択情報付与部を含む構成とし、選択情報をクラスタ特徴とともに表示するようにしてもよい。

【０１６２】

実施の形態５による文書分類装置を構成する情報処理システムは、図１に示したように実施の形態１と同様であるので、その説明は省略する。また、サーバー１０１およびクライアント１０２のハードウエア構成についても、図２・図３に示したように実施の形態１と同様であるので、その説明は省略する。

【０１６３】

つぎに、実施の形態５による文書分類装置の機能的構成について説明する。図１３は、この発明の実施の形態５による文書分類装置の構成を機能的に示すブロック図である。図１３において、実施の形態３の図９と同一のものに関しては同じ符号を付して、その説明を省略する。

【０１６４】

図１３のブロック図において、文書分類装置は、入力部４０１、言語解析部４０２、ベクトル生成部４０３、分類部４０４、分類パラメータ指示部４０５、分類結果記憶部４０６、クラスタ特徴表示部４０７、クラスタ特徴算出部４０８、分類体系記憶部４０９、クラスタ選択指示部４１０、分類体系閲覧操作部４１１、ベクトル記憶部９０１、文書表現空間修正部９０２のほか、選択情報付与部１３０１を含む構成である。

【０１６５】

選択情報付与部１３０１は、分類部４０４により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する。また、クラスタ特徴表示部４０７は、クラスタ特徴を表示するとともに、選択情報付与部１３０１により付与された選択情報を表示する。

【０１６６】

なお、選択情報付与部１３０１は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、機能を実現する。

【０１６７】

つぎに、選択情報付与部１３０１の詳細な内容について説明する。アンケートの調査の例において、独自性の高いユニークな意見は貴重であることが経験的に知られている。これは、調査を企画する担当者が予想できなかった意見である場合が多いからである。

【０１６８】

そこで、操作者に選択されたクラスタに所属する文書を、次回以降の分類実行

の際に使用する場合において、クラスタ特徴表示部４０７で個々の文書を表示する際に、各文書が何回選択されたかを示すことで、多重に利用される文書の識別性を向上させ、かつ一度も選択されない文書の識別性も向上させることができる。

【０１６９】

図１４は、実施の形態５による文書分類装置の分類結果記憶部４０６において設けられたテーブル１４００を示す説明図である。図１４において、文書ＩＤごとにテーブル化されており、テーブル１４００は、各文書がどのサイクルに分類実行の際に操作者に選択されたかを記録する。すなわち、選択された場合は選択情報として「１」を記録し、選択されなかった場合は選択情報として「０」を記録する。

【０１７０】

たとえば、４回分類が実行された際、文書ＩＤの「１」、第１回目および第２回目の分類実行時に操作者に選択されたことを示し、第３回目、第４回目の分類実行時には選択されなかったことを示している。一方、文書ＩＤの「２」は、未だ一度も選択されておらず、操作者にとって未知の意見という可能性を示唆している。

【０１７１】

こうした情報に基づいて、クラスタ特徴表示部４０７が文書を操作者に表示する際、たとえば、選択された回数に応じて表示を変化させるようにするとよい。変化させる視覚的特性としては、たとえば文字や背景の色の濃度や彩度等が考えられる。

【０１７２】

また、直接的に数字やグラフ等で選択された回数を表現することもできる。いずれにしてもよく選択される文書と一度も選択されていない文書とを視覚的に識別できる表示形式であれば、上記のものに限らない。

【０１７３】

また、上記選択情報を分類体系閲覧操作部４１１の閲覧操作により閲覧できるようにしてもよい。

【０１７４】

つぎに、選択情報付与部１３０１の処理の内容について説明する。図１５は、実施の形態５による文書分類装置の選択情報付与部１３０１の処理の手順を示すフローチャートである。図１５のフローチャートにおいて、まず、分類処理がおこなわれ（ステップＳ１５０１）、それに引き続き、最初の文書が抽出される（ステップＳ１５０２）。

【０１７５】

抽出された文書が、ステップＳ１５０１における分類処理の際に選択されたか否かを判断する（ステップＳ１５０３）。ここで、選択された場合（ステップＳ１５０３肯定）は、選択情報としてデータ「１」を記録する（ステップＳ１５０４）。一方、選択されなかった場合（ステップＳ１５０３否定）は、選択情報としてデータ「０」を記録する（ステップＳ１５０５）。

【０１７６】

つぎに、すべての文書について処理が終了したか否かを判断する（ステップＳ１５０６）。ここで、すべての文書について処理が終了していない場合（ステップＳ１５０６否定）は、つぎに文書を抽出し（ステップＳ１５０７）、ステップＳ１５０３へ移行し、以後、ステップＳ１５０３～Ｓ１５０７を繰り返しおこなう。

【０１７７】

一方、ステップＳ１５０６において、すべての文書について処理が終了した場合（ステップＳ１５０６肯定）は、ステップＳ１５０１へ移行し、再度分類処理がおこなわれる（ステップＳ１５０１）。このようにして、分類処理がおこなわれる回数だけ、ステップＳ１５０１～Ｓ１５０７の各処理が繰り返しおこなわれる。

【０１７８】

以上説明したように、実施の形態５によれば、選択情報付与部１３０１が選択情報を付与し、その選択情報をクラスタ特徴表示部４０７が表示するので、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができる。

【０１７９】

なお、実施の形態１～５で説明した文書分類方法は、あらかじめ用意されたプログラムをパーソナルコンピュータやワークステーション等のコンピュータで実行することにより実現される。このプログラムは、ハードディスク、フロッピーディスク、ＣＤ－ＲＯＭ、ＭＯ、ＤＶＤ等のコンピュータで読み取り可能な記録媒体に記録され、コンピュータによって記録媒体から読み出されることによって実行される。またこのプログラムは、上記記録媒体を介して、インターネット等のネットワークを介して配布することができる。

【０１８０】

【発明の効果】

以上説明したように、請求項１の発明によれば、入力手段が、文書データを入力し、言語解析手段が、前記入力手段により入力された文書データを解析して言語解析情報を得、ベクトル生成手段が、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類手段が、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出手段が、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、分類体系記憶手段が、前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶するので、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０１８１】

また、請求項２の発明によれば、入力手段が、文書データを入力し、言語解析手段が、前記入力手段により入力された文書データを解析して言語解析情報を得、ベクトル生成手段が、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類手段が、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類

し、文書の部分集合を生成し、クラスタ特徴算出手段が、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、表示手段が、前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示し、クラスタ選択指示手段が、前記表示手段に表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択し、分類体系記憶手段が、前記クラスタ選択指示手段により選択されたクラスタ特徴を分類体系の構成要素として記憶するので、選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０１８２】

また、請求項３の発明によれば、請求項２の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、ベクトル修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトルを前記分類手段により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正し、前記分類手段が、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類するので、既知になったクラスタの影響を排除した新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０１８３】

また、請求項４の発明によれば、請求項２の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、文書表現空間修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正し、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類するので、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分

類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０１８４】

また、請求項５の発明によれば、請求項３の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、文書表現空間修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正し、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類するので、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０１８５】

また、請求項６の発明によれば、請求項２または４の発明において、選択情報付与手段が、前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与し、前記表示手段が、前記クラスタ特徴を表示するとともに、選択情報付与手段により付与された選択情報を表示するので、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０１８６】

また、請求項７の発明によれば、請求項２～６の発明において、前記分類体系記憶手段が、前記選択指示手段により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは

一部および／または操作者が作成した情報を分類体系の構成要素として記憶するので、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できることので、分類体系の利用価値を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０１８７】

また、請求項８の発明によれば、入力工程が、文書データを入力し、言語解析工程が、前記入力工程により入力された文書データを解析して言語解析情報を得、ベクトル生成工程が、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出工程が、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、分類体系生成工程が、前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成するので、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０１８８】

また、請求項９の発明によれば、入力工程が、文書データを入力し、言語解析工程が、前記入力工程により入力された文書データを解析して言語解析情報を得、ベクトル生成工程が、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出工程が、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、表示工程が、前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示し、クラスタ選択指示工程が、前記表示工程に表示された複数のクラスタ特徴の中から所望のクラス

タ特徴を選択し、分類体系生成工程が、前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成するので、選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０１８９】

また、請求項１０の発明によれば、請求項９の発明において、ベクトル修正工程が、前記ベクトル生成工程により生成された文書特徴ベクトルを前記分類工程により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正し、前記分類工程が、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類するので、既知になったクラスタの影響を排除した新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０１９０】

また、請求項１１の発明によれば、請求項９の発明において、文書表現空間修正工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正し、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類するので、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０１９１】

また、請求項１２の発明によれば、請求項１０の発明において、文書表現空間修正工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似

度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて修正し、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類するので、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０１９２】

また、請求項１３の発明によれば、請求項９または１１の発明において、選択情報付与工程が、前記分類工程により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与し、前記表示工程が、前記クラスタ特徴を表示するとともに、選択情報付与工程により付与された選択情報を表示するので、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０１９３】

また、請求項１４の発明によれば、請求項９～１３の発明において、前記分類体系生成工程が、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成するので、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できることので、分類体系の利用価値を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０１９４】

また、請求項１５の発明によれば、請求項８～１４のいずれか一つに記載され

た方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項８～１４の動作をコンピュータによって実現することが可能な記録媒体が得られるという効果を奏する。

【図面の簡単な説明】

【図１】

この発明の実施の形態１による文書分類装置を構成する情報処理システム全体のハードウエア構成を示す説明図である。

【図２】

実施の形態１による文書分類装置を構成する情報処理システムにおけるサーバーをハードウエア的に示す説明図である。

【図３】

実施の形態１による文書分類装置を構成する情報処理システムにおけるクライアントをハードウエア的に示す説明図である。

【図４】

実施の形態１による文書分類装置の構成を機能的に示すブロック図である。

【図５】

実施の形態１による文書分類装置のクラスタ特徴表示部の表示の一例を示す説明図である。

【図６】

実施の形態１による文書分類装置の一連の処理の手順を示すフローチャートである。

【図７】

この発明の実施の形態２による文書分類装置の構成を機能的に示すブロック図である。

【図８】

実施の形態２よる文書分類装置の一連の処理の手順を示すフローチャートである。

【図９】

この発明の実施の形態３による文書分類装置の構成を機能的に示すブロック図

である。

【図１０】

実施の形態３よる文書分類装置の一連の処理の手順を示すフローチャートである。

【図１１】

この発明の実施の形態４による文書分類装置の構成を機能的に示すブロック図である。

【図１２】

実施の形態４よる文書分類装置の一連の処理の手順を示すフローチャートである。

【図１３】

この発明の実施の形態５による文書分類装置の構成を機能的に示すブロック図である。

【図１４】

実施の形態５による文書分類装置の分類結果記憶部において設けられたテーブルを示す説明図である。

【図１５】

実施の形態５による文書分類装置の選択情報付与部の処理の手順を示すフローチャートである。

【符号の説明】

１０１　サーバー

１０２　クライアント

１０３　ネットワーク

２００　バス

２０１　ＣＰＵ

２０２　ＲＯＭ

２０３　ＲＡＭ

２０４　Ｉ／Ｆ

２０５　通信回線

２０６　ディスク装置

２０７　共有フォルダ

２０８　ディスプレイ

２０９　キーボード

２１０　マウス

３０１　ＣＰＵ

３０２　ＲＯＭ

３０３　ＲＡＭ

３０６　ハードディスク

３０８　ディスプレイ

３０９　Ｉ／Ｆ

３１１　キーボード

３１２　マウス

３１３　スキャナ

３１４　プリンタ

４０１　入力部

４０２　言語解析部

４０３　ベクトル生成部

４０４　分類部

４０５　分類パラメータ指示部

４０６　分類結果記憶部

４０７　クラスタ特徴表示部

４０８　クラスタ特徴算出部

４０９　分類体系記憶部

４１０　クラスタ選択指示部

４１１　分類体系閲覧操作部

５１０　カーソル

７０１，９０１，１１０１　ベクトル記憶部

７０２，１１０２　ベクトル修正部

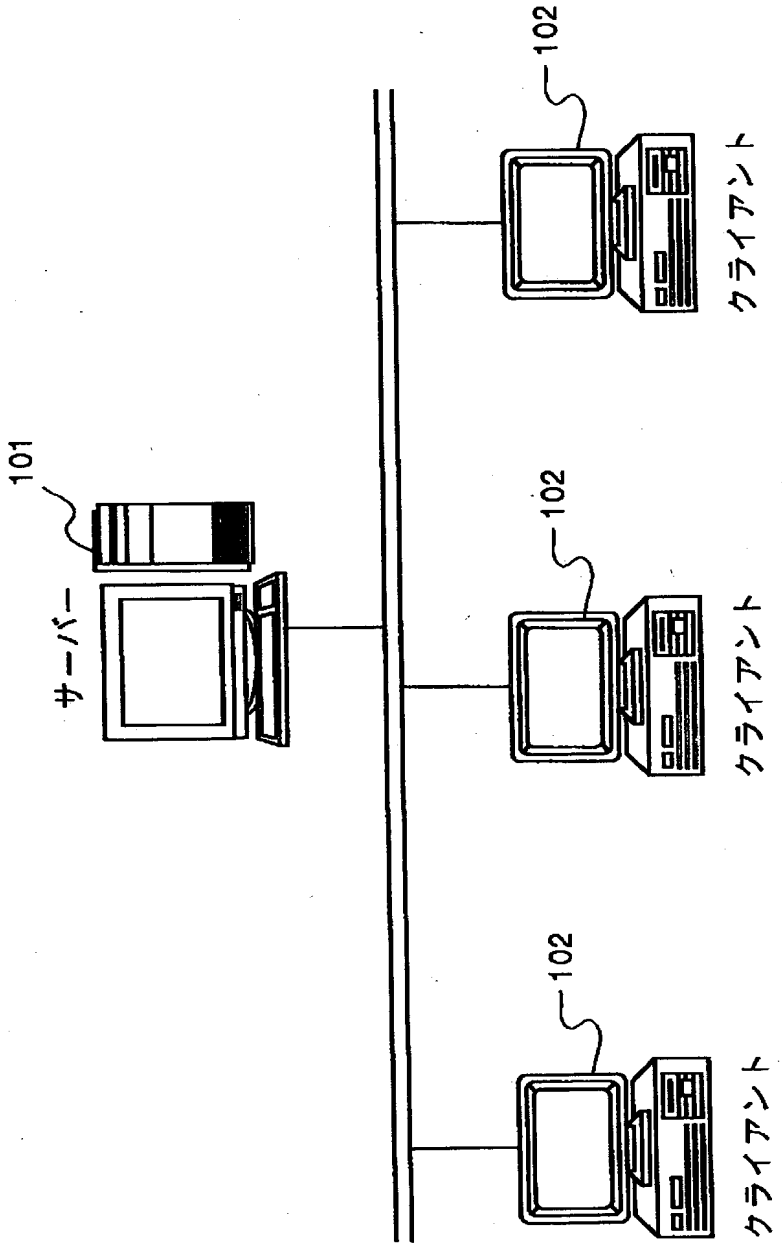
９０２，１１０３　文書表現空間修正部

１３０１　選択情報付与部

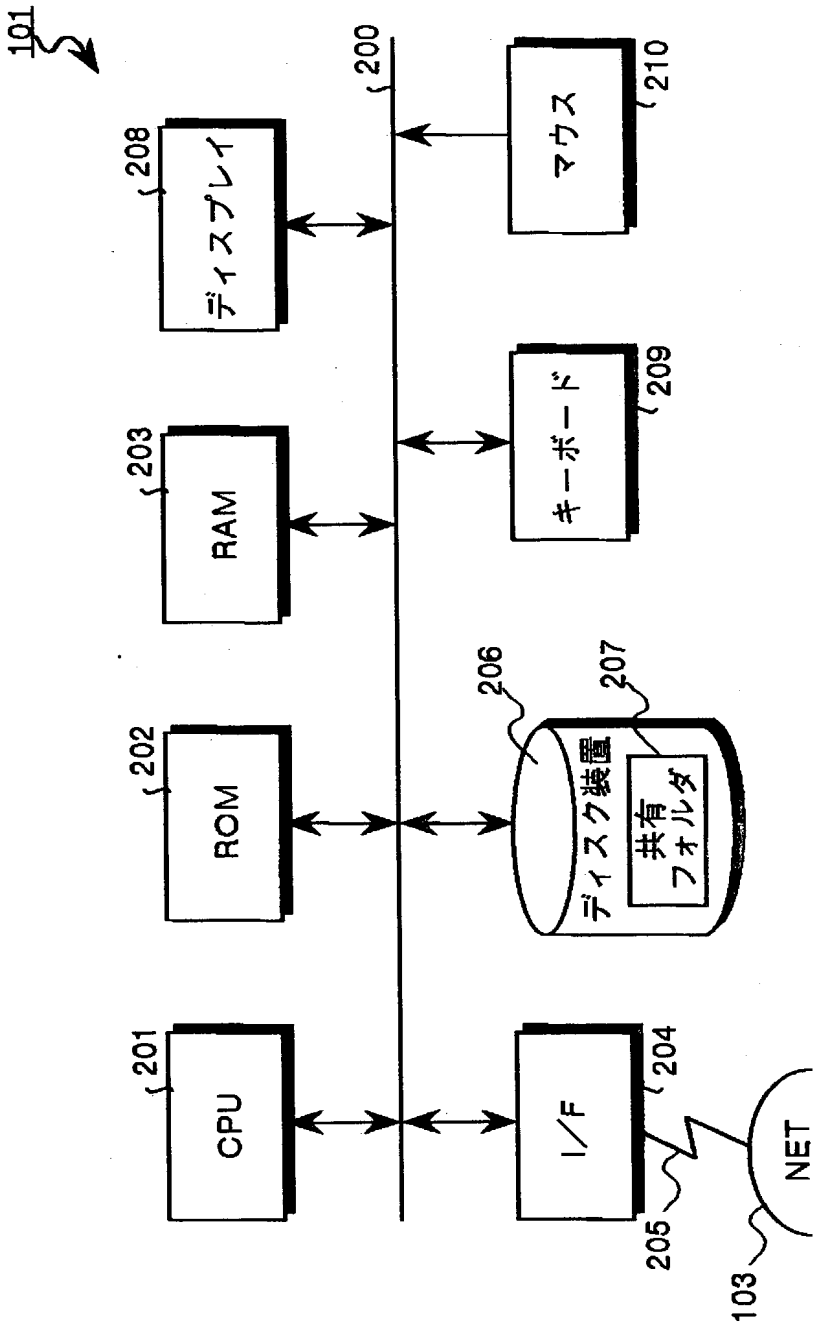
１４００　テーブル

【書類名】　図面

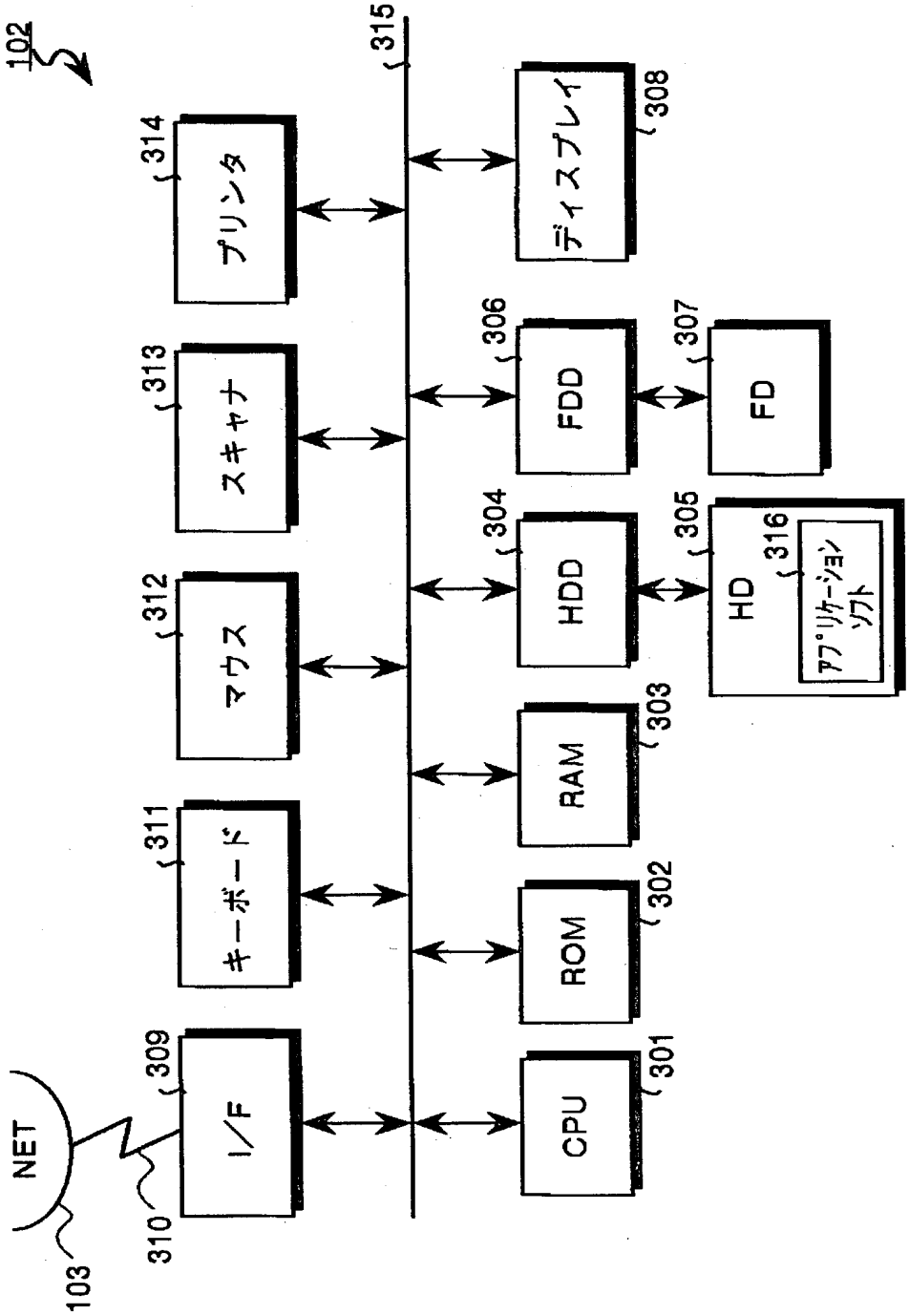
【図１】



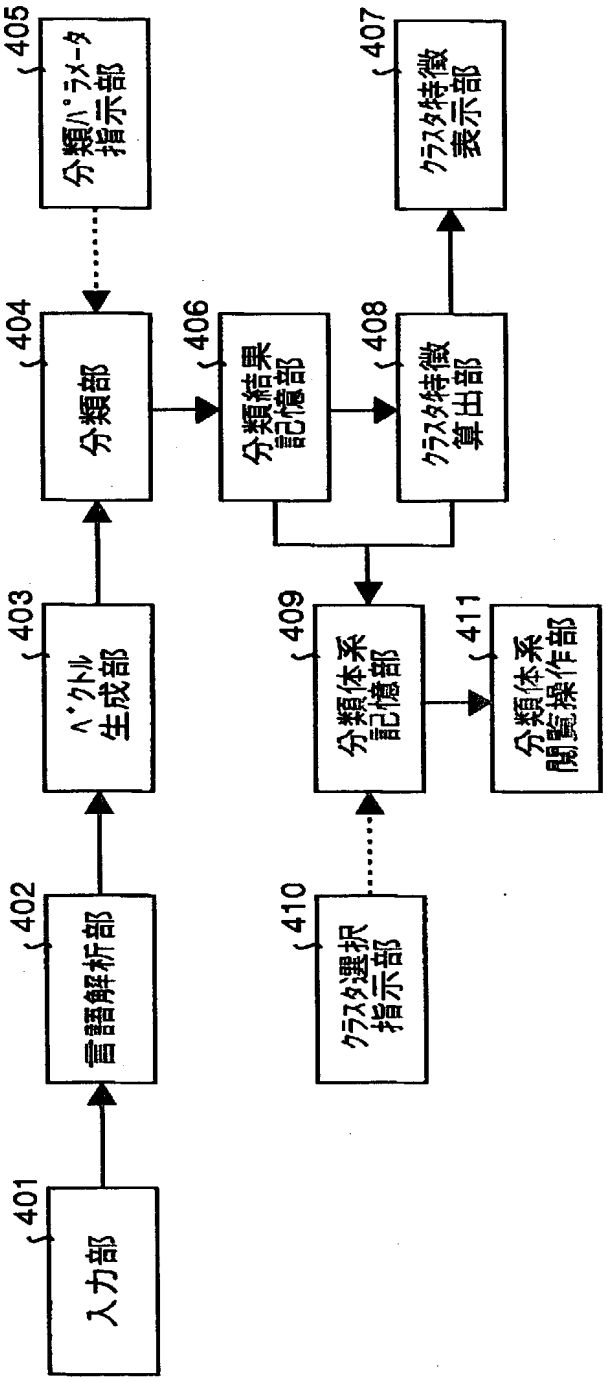
【図２】


101
200
208
ディスプレイ
210
マウス
209
キーボード
203
RAM
206
207
202
ROM
ディスク装置
共有
フォルダ
201
CPU
I／F
204
205
NET
103

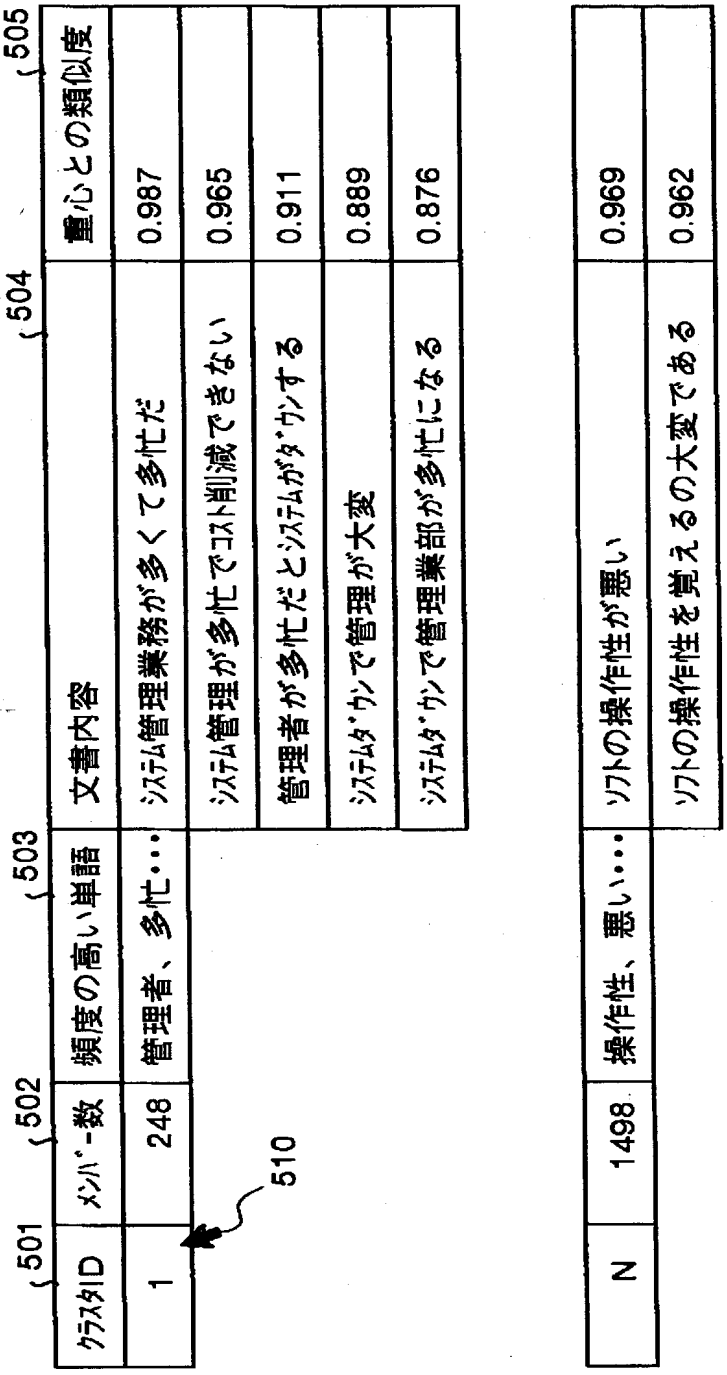
【図３】


102
315
314
プリンタ
308
ディスプレイ
313
スキャナ
306
FDD
307
FD
312
マウス
304
HDD
305
HD
316
アプリケーション
ソフト
311
キーボード
303
RAM
309
I/F
302
ROM
NET
310
301
CPU
103

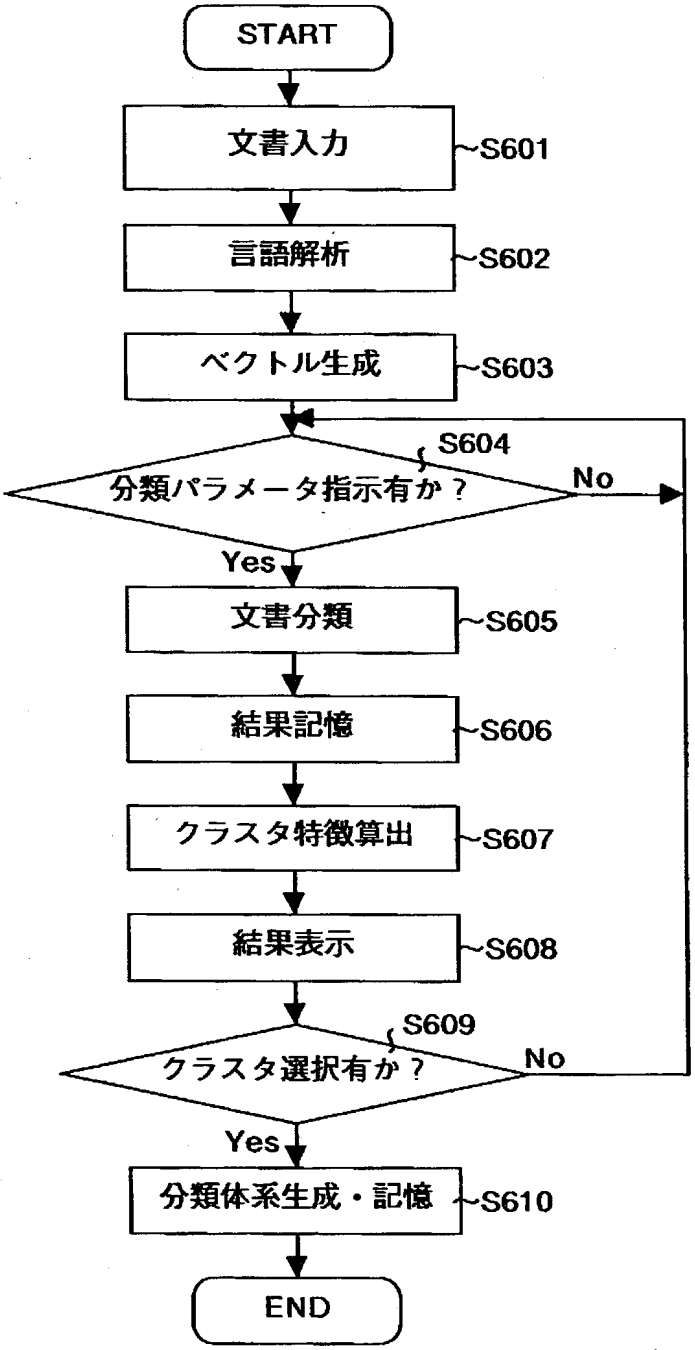
【図４】


401
入力部
402
言語解析部
403
ベクトル生成部
404
分類部
405
分類パラメータ指示部
406
分類結果記憶部
407
クラスタ特徴表示部
408
クラスタ特徴算出部
409
分類体系記憶部
410
クラスタ選択指示部
411
分類体系閲覧操作部

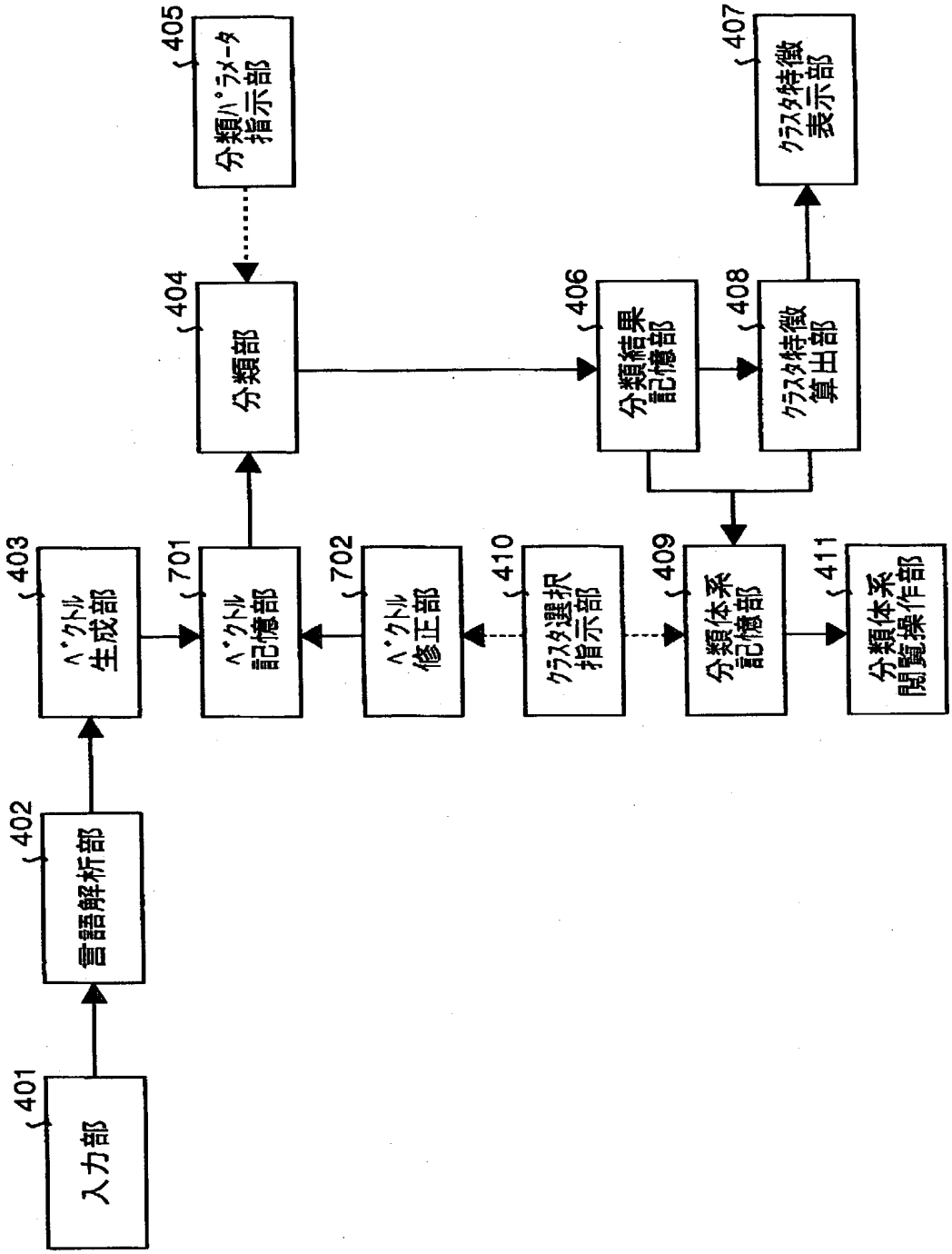
【図５】


501
502
503
504
505
510
クラスタID
メンバー数
頻度の高い単語
文書内容
重心との類似度
1
248
管理者、多忙・・・
システム管理業務が多くて多忙だ
0.987
システム管理が多忙でコスト削減できない
0.965
管理者が多忙だとシステムがダウンする
0.911
システムダウンで管理が大変
0.889
システムダウンで管理業部が多忙になる
0.876
N
1498
操作性、悪い・・・
ソフトの操作性が悪い
0.969
ソフトの操作性を覚えるの大変である
0.962

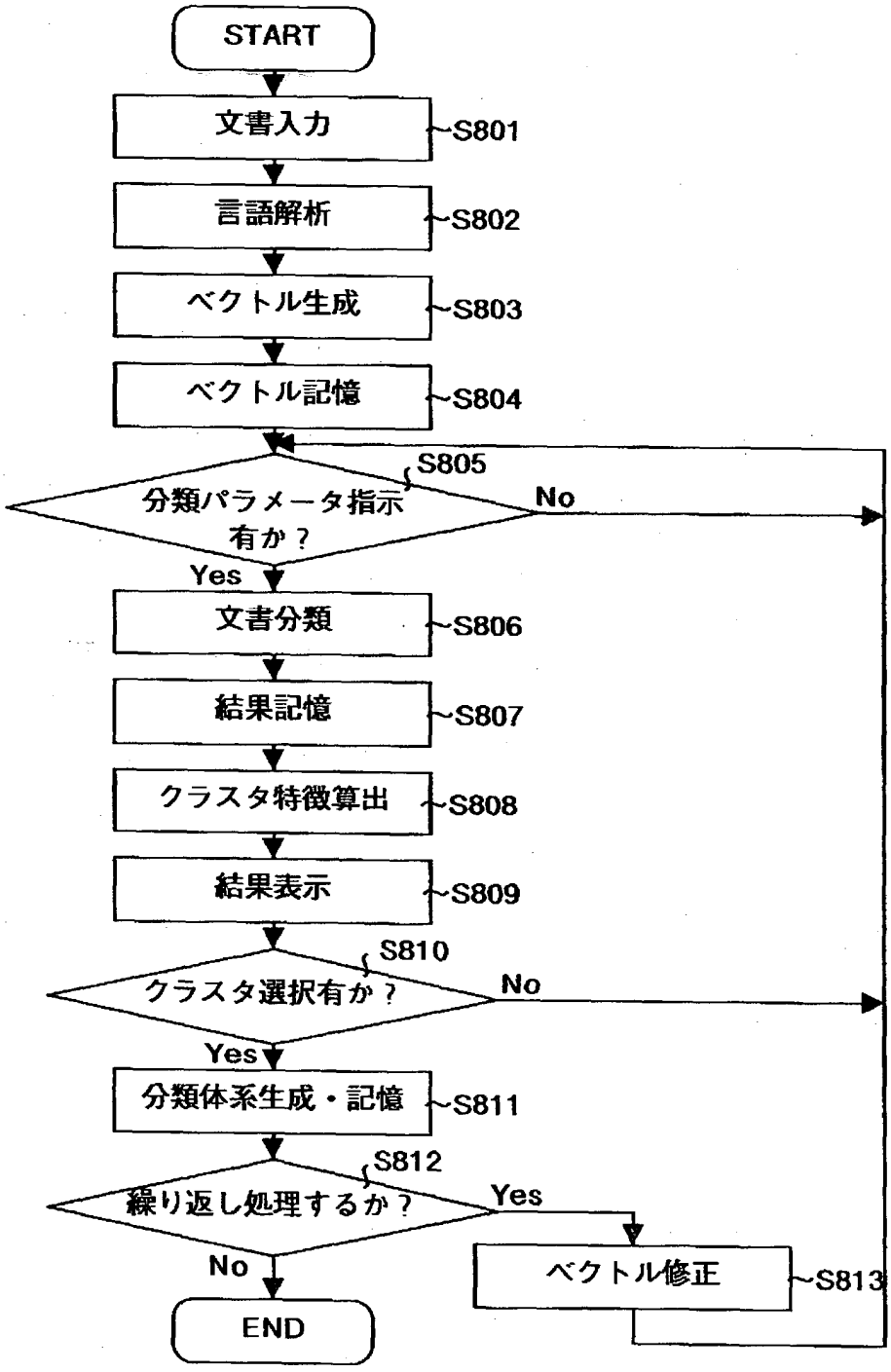
【図６】


START
文書入力
S601
言語解析
S602
ベクトル生成
S603
S604
分類パラメータ指示有か？
No
Yes
文書分類
S605
結果記憶
S606
クラスタ特徴算出
S607
結果表示
S608
S609
クラスタ選択有か？
No
Yes
分類体系生成・記憶
S610
END

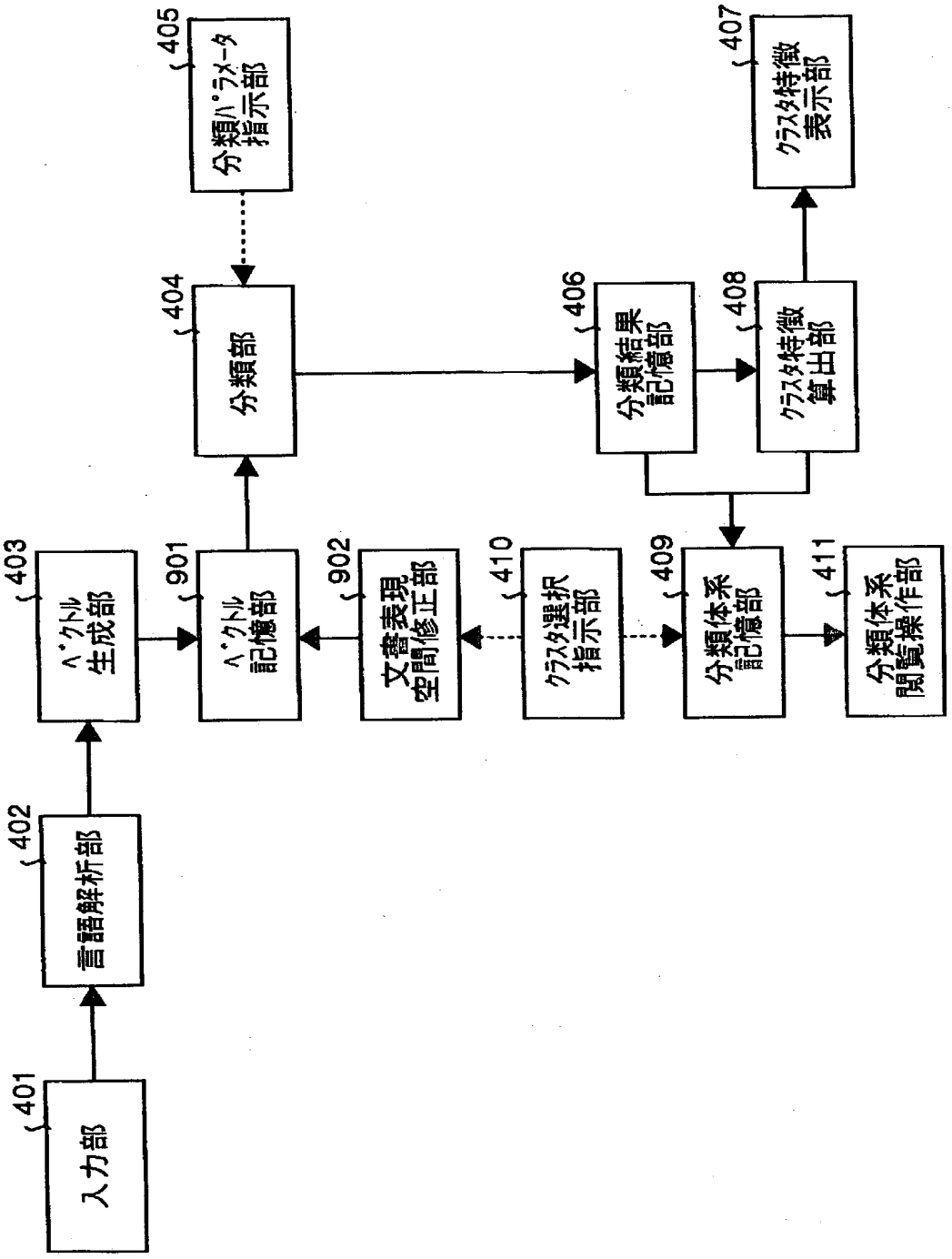
【図７】


401
入力部
402
言語解析部
403
ﾍﾞｸﾄﾙ生成部
701
ﾍﾞｸﾄﾙ記憶部
702
ﾍﾞｸﾄﾙ修正部
404
分類部
405
分類ﾊﾟﾗﾒｰﾀ指示部
406
分類結果記憶部
407
ｸﾗｽﾀ特徴表示部
408
ｸﾗｽﾀ特徴算出部
409
分類体系記憶部
410
ｸﾗｽﾀ選択指示部
411
分類体系閲覧操作部

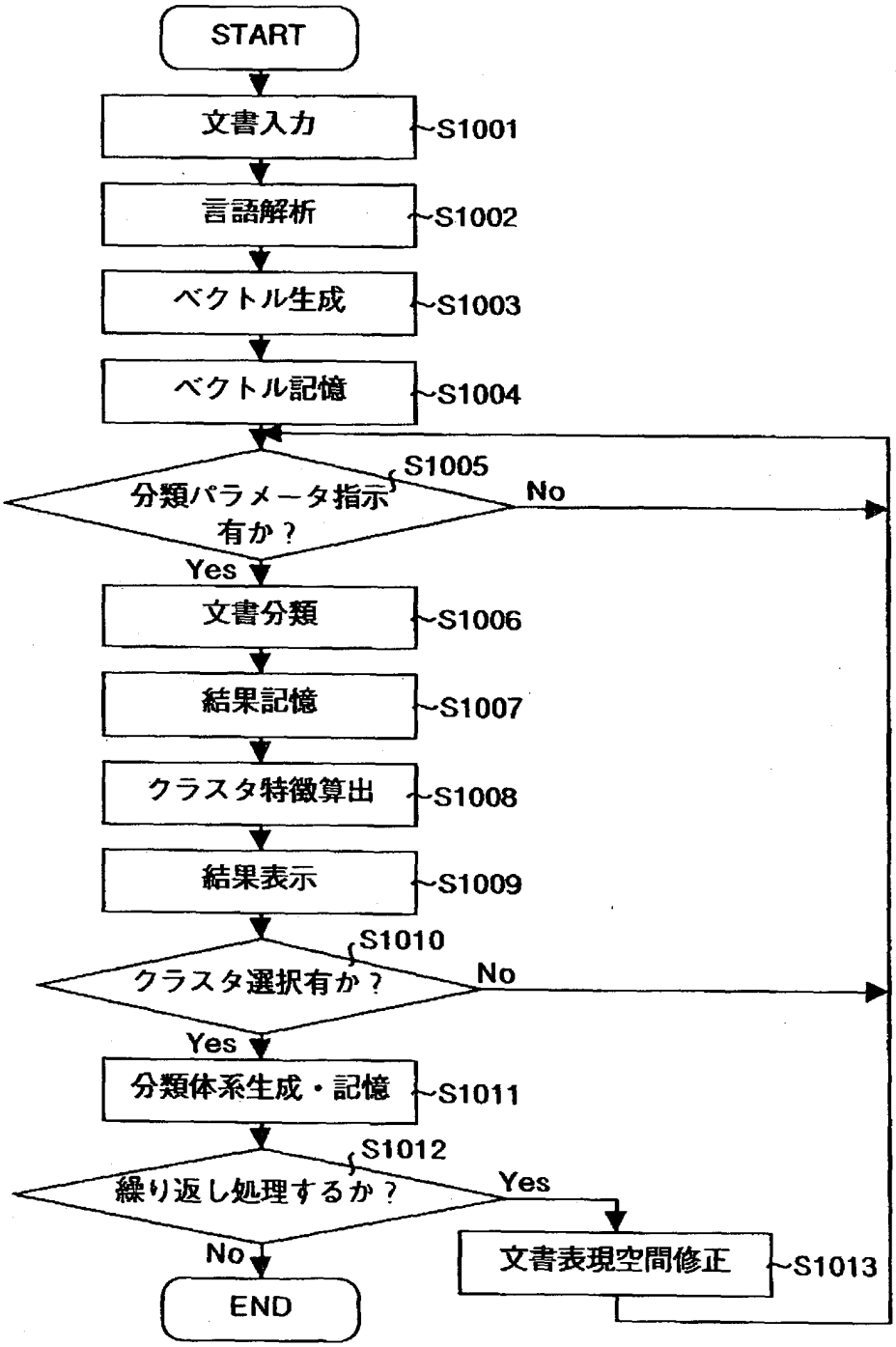
【図８】


START
文書入力
S801
言語解析
S802
ベクトル生成
S803
ベクトル記憶
S804
S805
分類パラメータ指示有か？
No
Yes
文書分類
S806
結果記憶
S807
クラスタ特徴算出
S808
結果表示
S809
S810
クラスタ選択有か？
No
Yes
分類体系生成・記憶
S811
S812
繰り返し処理するか？
Yes
No
ベクトル修正
S813
END

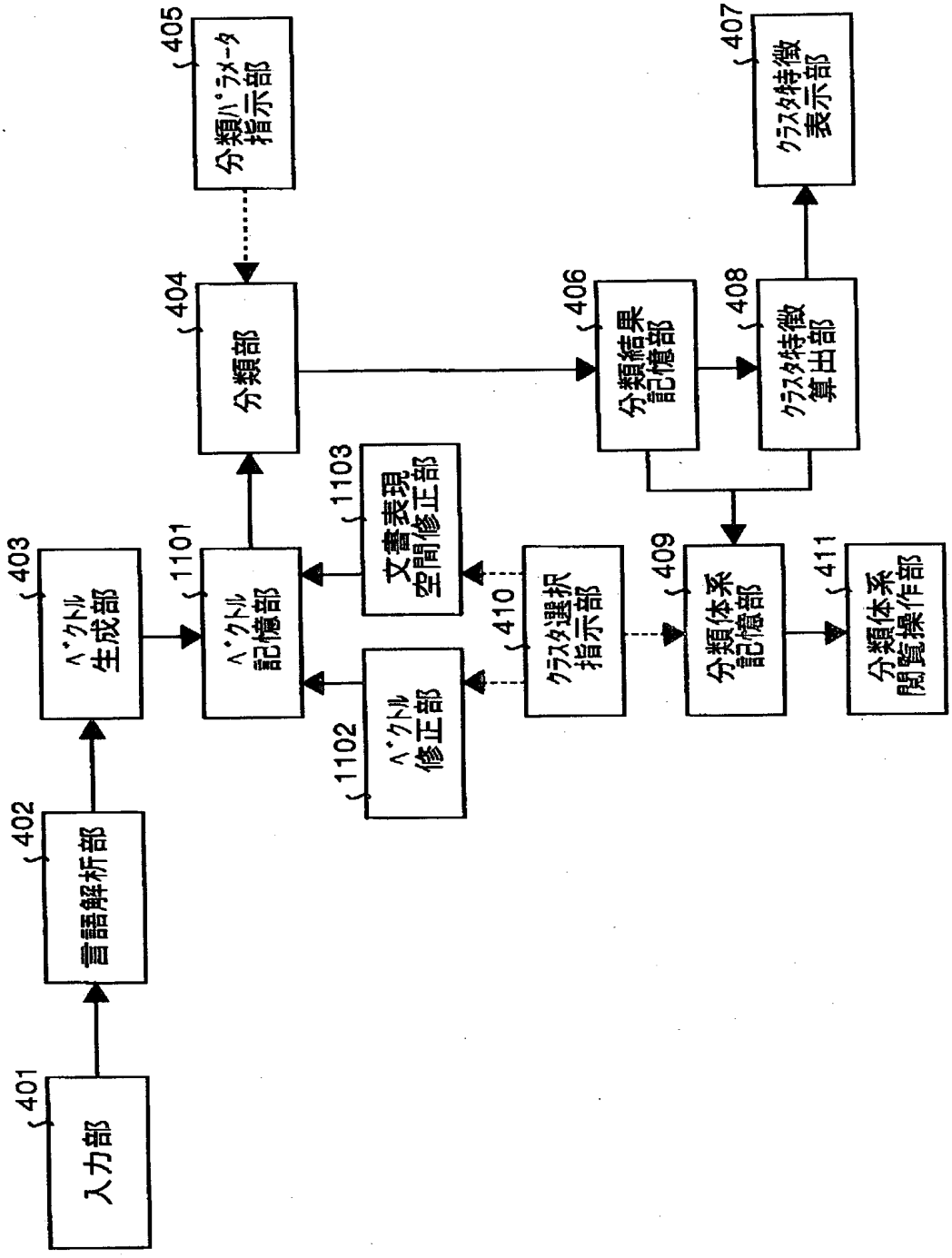
【図９】


401
入力部
402
言語解析部
403
ﾍﾞｸﾄﾙ生成部
901
ﾍﾞｸﾄﾙ記憶部
902
文書表現空間修正部
404
分類部
405
分類ﾊﾟﾗﾒｰﾀ指示部
406
分類結果記憶部
407
クラスタ特徴表示部
408
クラスタ特徴算出部
410
クラスタ選択指示部
409
分類体系記憶部
411
分類体系閲覧操作部

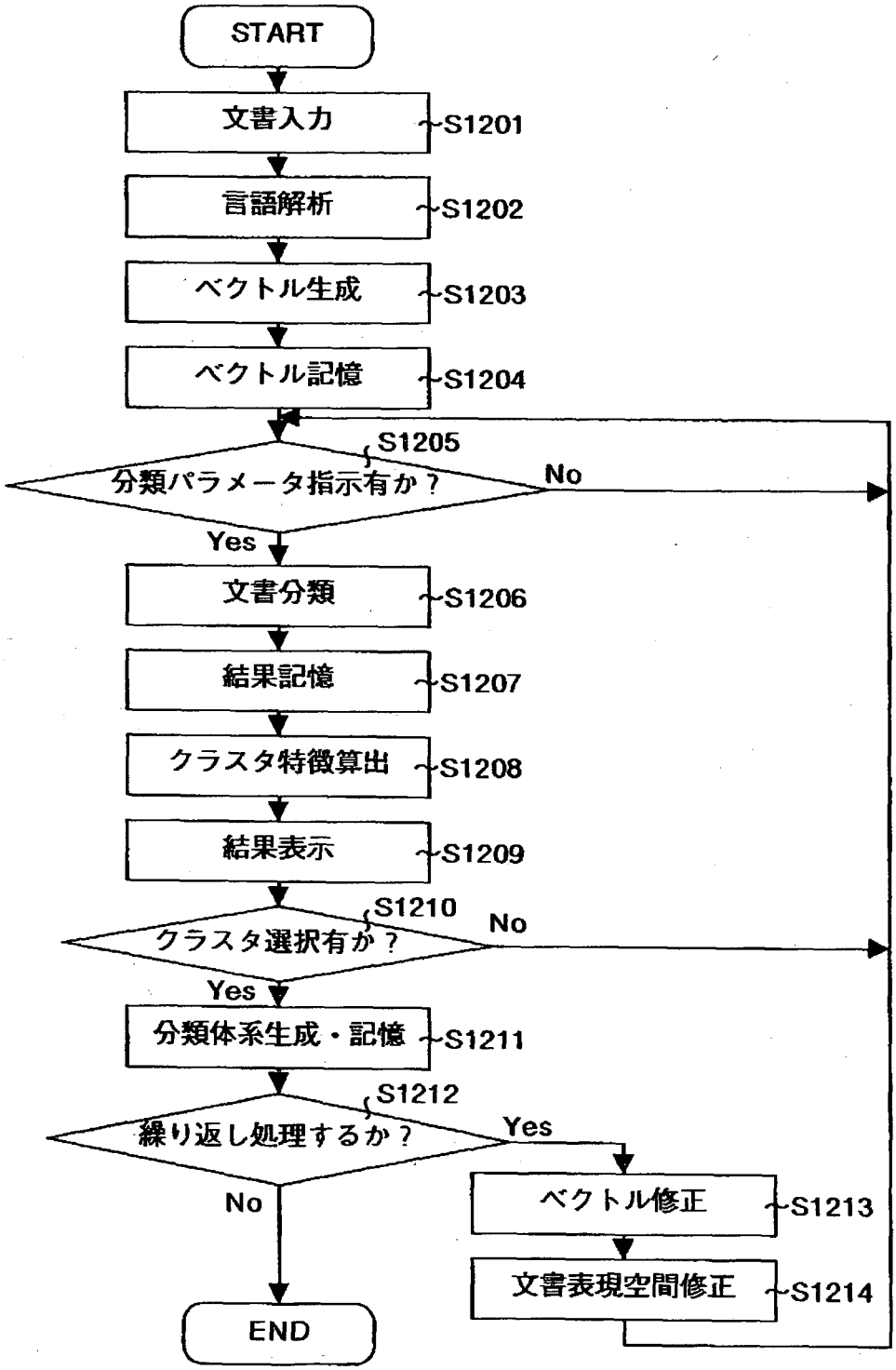
【図１０】


START
文書入力
S1001
言語解析
S1002
ベクトル生成
S1003
ベクトル記憶
S1004
S1005
分類パラメータ指示有か？
No
Yes
文書分類
S1006
結果記憶
S1007
クラスタ特徴算出
S1008
結果表示
S1009
S1010
クラスタ選択有か？
No
Yes
分類体系生成・記憶
S1011
S1012
繰り返し処理するか？
Yes
No
文書表現空間修正
S1013
END

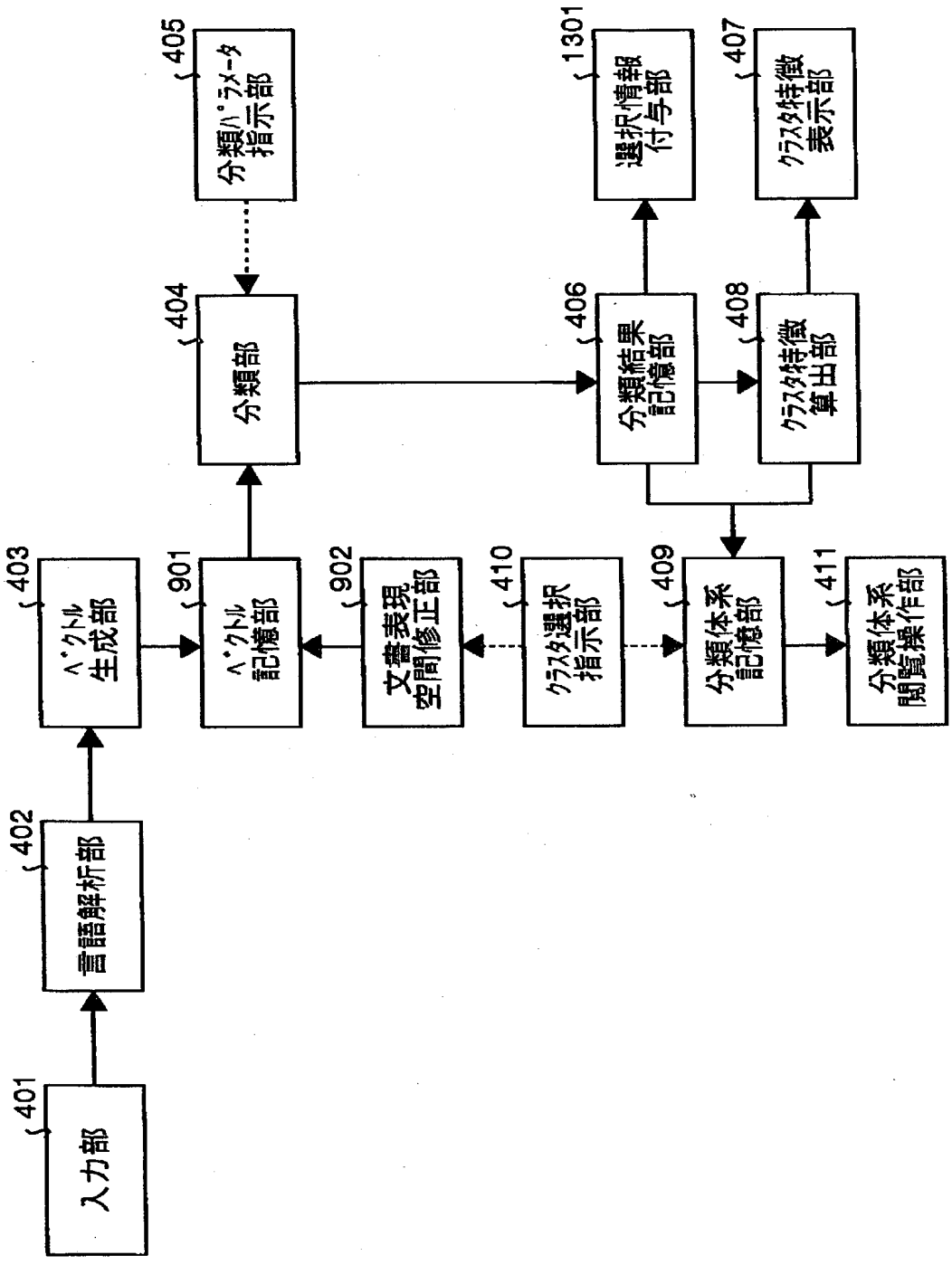
【図１１】


401
入力部
402
言語解析部
403
ﾍﾞｸﾄﾙ生成部
1101
ﾍﾞｸﾄﾙ記憶部
1102
ﾍﾞｸﾄﾙ修正部
1103
文書表現空間修正部
404
分類部
405
分類ﾊﾟﾗﾒｰﾀ指示部
406
分類結果記憶部
407
クラスタ特徴表示部
408
クラスタ特徴算出部
409
分類体系記憶部
410
クラスタ選択指示部
411
分類体系閲覧操作部

【図１２】


START
文書入力
S1201
言語解析
S1202
ベクトル生成
S1203
ベクトル記憶
S1204
S1205
分類パラメータ指示有か？
No
Yes
文書分類
S1206
結果記憶
S1207
クラスタ特徴算出
S1208
結果表示
S1209
S1210
クラスタ選択有か？
No
Yes
分類体系生成・記憶
S1211
S1212
繰り返し処理するか？
Yes
No
ベクトル修正
S1213
文書表現空間修正
S1214
END

【図１３】

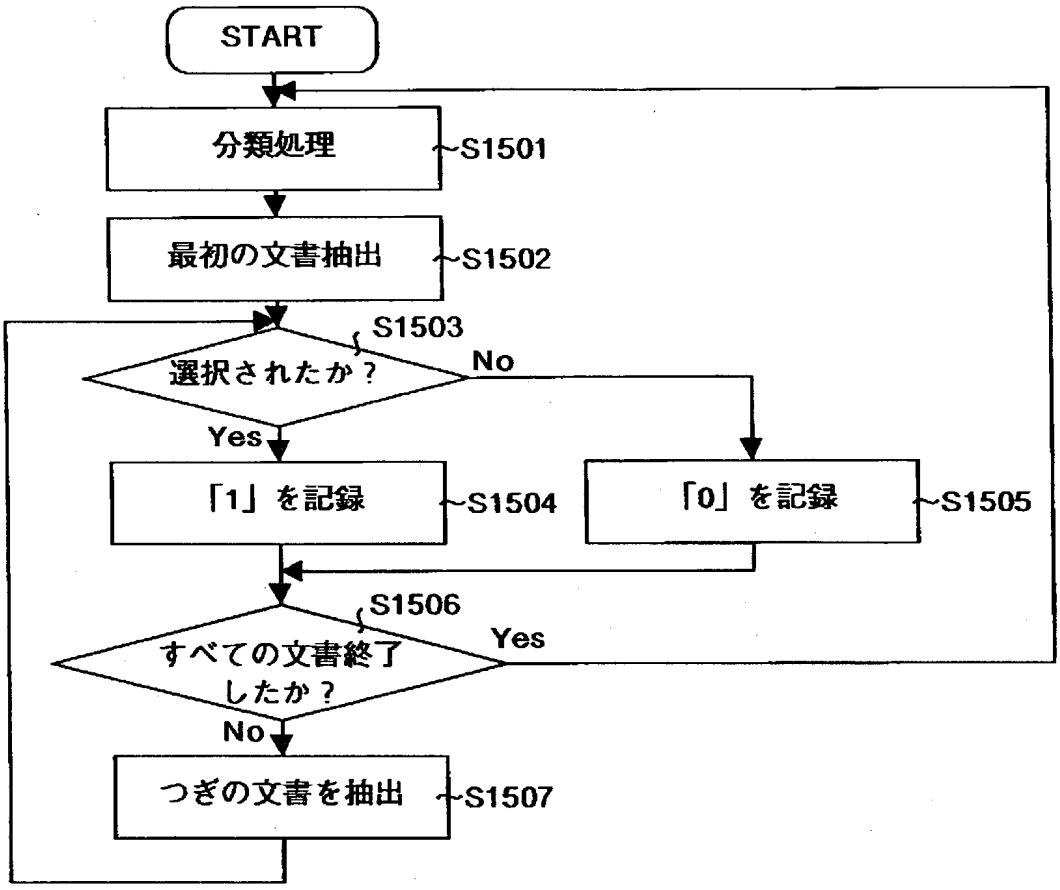

401
入力部
402
言語解析部
403
ﾍﾞｸﾄﾙ生成部
901
ﾍﾞｸﾄﾙ記憶部
902
文書表現空間修正部
410
クラスタ選択指示部
409
分類体系記憶部
411
分類体系閲覧操作部
404
分類部
405
分類ﾊﾟﾗﾒｰﾀ指示部
406
分類結果記憶部
408
クラスタ特徴算出部
1301
選択情報付与部
407
クラスタ特徴表示部

【図１４】

1400

文書ID	選択情報（１＝選択/０＝未選択）分類回数順
1	[1, 1, 0, 0]
2	[0, 0, 0, 0]
3	[0, 1, 0, 0]
n-1	[1, 0, 0, 0]
n	[0, 0, 0, 0]

【図１５】


START
分類処理
S1501
最初の文書抽出
S1502
S1503
選択されたか？
No
Yes
「1」を記録
S1504
「0」を記録
S1505
S1506
すべての文書終了したか？
Yes
No
つぎの文書を抽出
S1507

【書類名】　　要約書

【要約】

【課題】　文書の内容に基づいて文書の分類をおこなう文書分類装置において、任意の文書集合にどのような内容が含まれるかを漸次的に収集することを課題とする。

【解決手段】　文書データを入力する入力部４０１と、入力された文書データを解析して言語解析情報を得る言語解析部４０２と、得られた言語解析情報に基づいて文書データに対する文書特徴ベクトルを生成するベクトル生成部４０３と、生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類部４０４と、生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出部４０８と、算出されたクラスタ特徴を表示する表クラスタ特徴表示部４０７と、表示された複数のクラスタ特徴の中から所望のクラスタ特徴を選択するクラスタ選択指示部４１０と、選択されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶部４０９とを備える。

【選択図】　　図４

出 願 人 履 歴 情 報

識別番号　〔０００００６７４７〕

１．変更年月日　１９９０年　８月２４日
［変更理由］　新規登録
住　所　東京都大田区中馬込１丁目３番６号
氏　名　株式会社リコー